Using NMR to Identify Structural Features of Lin28-Regulated miRNAs and mRNAs and as a Tool for Comparing Differences in Cellular Metabolism

Using NMR to identify structural features of Lin28-regulated miRNAs and mRNAs and as a tool for comparing differences in cellular metabolism

---

A dissertation presented

by

Elizabeth Mary O'Day

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

May 2013

Dissertation Advisors: Dr. Judy Lieberman & Dr. Gerhard Wagner

Elizabeth Mary O'Day

Using NMR to identify structural features of Lin28 regulated miRNAs and mRNAs and as a tool for comparing differences in cellular metabolism

Abstract

Part 1 of this thesis seeks to identify shared structural features of Lin28-regulated miRNAs and mRNAs. Lin28 is an evolutionarily conserved, RNA binding protein, highly expressed in stem cells and poorly differentiated cancers, that inhibits differentiation and helps maintain stem cell properties. Lin28 binds to both the loops of let-7 precursors to block let-7 biogenesis and to Lin28 responsive elements (LREs) in mRNAs either to enhance or inhibit translation. Lin28 RNA binding properties are not well defined. We used NMR spectroscopy, fluorescence assays and bioinformatics to identify common features of Lin28 targets. We show that Lin28 binds G-rich sequences that have properties of G-quartets (G4s). Based on mutational analysis, we show that G4s are important for Lin28 binding. Upon binding, Lin28 may unwind the G4 structure. Our findings suggest that Lin28 recognizes G-quartets in the RNAs it regulates and might function to unwind them.

In part 2 of this thesis we use an unbiased NMR metabolite screening method to identify glucose metabolites differentially produced in BPLER and HMLER isogenic triple negative breast cancer cell lines that have dramatic differences in tumor initiating capacity. N-acetylneuraminic acid (Neu5Ac), a

sugar added to the end of glycosylation chains, is much more abundant in BPLER than HMLER cells. Manipulating Neu5Ac expression using neuraminidase or siRNA knockdown of the Neu5Ac biosynthetic enzymes N-acetylneuraminic acid synthase (NANS) or cytidine monophosphate N-acetylneuraminic acid synthase (CMAS), reduces *in vitro* invasivity of BPLER cells. CMAS protein is also increased in BPLER relative to HMLER. Overexpressing CMAS in HMLER cells increases their invasiveness. Moreover, stable knockdown of CMAS blocks BPLER tumor growth in xenografted mice. Thus increased Neu5Ac synthesis is linked to tumor initiation and invasivity in a human triple negative breast cancer cell line.

Table of Contents

Acknowledgments

First, I would like to thank my dissertation advisors Dr. Judy Lieberman and Dr. Gerhard Wagner. For my PhD I had the rare opportunity to work in two laboratories. While no doubt difficult, I am beyond grateful for this opportunity. I have been exposed to more science than I thought possible. I have also had the opportunity to work on numerous projects and gain experience with many different techniques. As I continue in science I feel confident I can draw on lessons I've learned in each laboratory. Thank you.

Judy has been an incredible mentor. I am truly amazed at the breadth of science she has mastered. Not only has she taught me how to approach the scientific process, but she's also given me confidence in my abilities to think creatively about a problem. As a woman in science her courage and willingness to fight for what she believes in are qualities to which I aspire. I will forever cherish our discussions on all the "wish we had the time to do projects". Thank you for everything.

As an NMR spectroscopist there is no better person in the world to work for than Gerhard Wagner. His intelligence is unmatched and he is by far one of the kindest people I have ever met. Gerhard's careful insight and never-ending curiosity has shaped how I do science. He has also taught me that no matter what the data says, it's important to share a few laughs. Gerhard has supported me on every project science and personal. Thank you for always believing in me.

I have also been incredibly fortunate to work with talented and inspiring colleagues. Thank you Hari, Ricard, Vlado, Mayaan, Dharma, Pat, Kendra, Greg

H., Greg I., Ricardo, Radiana, Fabio, Michael, Mynn, Minh, Danielle, Marshall, Nan, Ashish and Zhan. They've become an extended family to me and I would not have survived without them (literally – thank you for reminding me to eat). Additionally, thank you to all past and present members of the Lieberman and Wagner labs that made conducting research an exciting and enjoyable experience.

I would also like to thank my Dissertation Advisory Committee including Gary Ruvkun, Jack Szostak and Victoria D'Souza. Their support, guidance and helpful suggestions have been critical in the development of this project. Thank you to my defense committee members Anna Krichevsky, Hao Wu and Victor Ambros for their participation in my defense. I'd also like to offer special thanks to the BBS program that welcomed me into the BBS family.

Finally I would like to thank my parents, brothers, grandmother, aunts, uncles, cousins and friends, especially Shannon, Kerry, Sarah, Marissa, Q, Alissa, JW, DR, co-capt, Duffy, POB, Dane, Andy and TD. Your unswerving belief in my gives me the strength to pursue the limits of my dreams. Thank you for your constant support. I love you more than words can describe.

.

Abbreviations

NMR spectroscopy: nuclear magnetic resonance spectroscopy

L/A switch: larva-to-adult switch

CSD: cold shock domain

ZnF: zinc finger motif

GRP2: glycine rich protein 2

ESC: embryonic stem cell

iPSc: induced pluripotent stem cell

GWAS: genome wide association study

SNP: single nucleotide polymorphism

miRNA: microRNA

nt: nucleotide

pri-miRNA: primary microRNA

pre-miRNA: precursor microRNA

RISC: RNA-induced-silencing-complex

TUT4: Tutase4

LRE: Lin28 responsive element

UTR: untranslated region

RHA: RNA helicase A

IP: immunoprecipitation

EMSA: electrophoretic mobility shift analysis

G-rich: guanine-rich

G4: G-quartet

NOESY: Nuclear Overhauser Effect Spectroscopy

PAGE: polyacrylamide gel electrophoresis

NMM: N-methyl mesoporphyrin IX

UG4U: UGGGGU

FMRP: Fragile-X-mental-retardation-protein-1

K-S test: Kolmogorov-Smirnov test

TBA: thrombin binding aptamer

cel-pre-let-7: *C. elegans* pre-let-7

IPTG: isopropyl D-1-thiogalactopyranoside

BME: β-mercaptoethanol

Neu5Ac: N-acetylneuraminic acid

NANS: N-acetylneuraminic acid synthase

CMAS: cytidine monophosphate N-acetylneuraminic acid synthetase

TNBC: triple negative breast cancer

DSS: 4,4-dimethyl-4-silapentane-1-sulfonic acid

HSQC: heteronuclear single quantum coherence

HMBD: Human Metabolome database

LC-M/S: liquid chromatography mass spectrometry

MRM: multiple reaction monitoring

UDP-GlcNAc: UDP-N-acetylglucosamine

GNE: glucosamine (UDP-N-acetyl)-2-epimerase

NANP: N-acetylneuraminic acid phosphatase

CMP-Neu5Ac: cytidine-5-monophosphate N-acetylneuraminic acid

ST: sialyltransferase

2-KB: 2-ketobutryic acid

HK1: hexokinase 1

GFA: glucosamine-fructose-6-phosphate aminotransferase

GlcN-6-P: glucosamine-6-phosphate

GNA1: glucosamine-6-P-N-acetyltransferase

GlcNAc-6-P: N-acetyglucosamine-6-phosphate

PGM3: phoso-acetyl-glucosamine mutase

GlcNAc-1-P: N-acetylglucosamine-1-phosphate

UAP: UDP-N-acetylglucosamine pyrophosphorylase

ManNAc: N-acetyl-D-mannosamine

ManNAc-6-P: N-acetyl-D-mannosamine-6-phosphate

PEP: phosphoenolpyruvate

Neu5Ac-9-P: N-acetylneuraminic acid-9-phosphate

WGA: wheat germ agglutinin

BPLER-shCMAS: CMAS stably knocked down BPLER cells

BPLER-EV: Empty vector shRNA BPLER cells

List of Tables & Figures

Chapter 1

Chapter 2

Dedication

To my grandmother, Margaret Patricia Riley O'Day, at 93 years of age your beauty and intelligence continue to inspire me.

Attributions

In part 1, Dr. Mynn Tan calculated G-scores for Lin28 mRNA datasets (**Figure 2.13**). Part 2 of this thesis was done in close collaboration with Dr. Greg Idos, a postdoctoral fellow in the Lieberman lab. I assisted Dr. Idos with invasion assays (**Figures 4.10, 4.13-4.14**). Dr. Idos performed the *in vivo* tumor studies (**Figure 4.16**). I performed the mass spectrometry experiments with Colin Hill, a senior research associate at Agios Pharmaceuticals (**Figure 4.7**).

Part 1: Identifying common structural features of Lin28-regulated miRNAs and mRNAs

Chapter 1: Introduction to Lin28

The Lin28 protein family regulates networks of RNAs that control developmental timing (*1-3*), pluripotency (*4, 5*), metabolism (*6, 7*), and tumorigenesis (*8-10*). Lin28 directly binds to the precursor miRNAs (pre-miRNAs) of the let-7 family of miRNAs and to mRNAs that function in maintaining pluripotency and regulating cell cycle progression (*11-15*). How Lin28 selects RNA targets remains unclear. Properties of Lin28 binding RNAs have been proposed based on mutagenesis analysis, structural studies, molecular modeling and sequencing of Lin28-bound RNAs. Uncovering the mechanism by which Lin28 selects RNA targets and regulates them is critical to understanding the function of Lin28.

The work in part 1 of this thesis seeks to identify shared structural features of Lin28-binding miRNAs and mRNAs that contribute to Lin28 recognition. I will first review what is known about the biological role of Lin28 and the identity of its targets. I will then discuss studies to identify the RNA determinants that interact with Lin28.

The function of Lin28

Lin28 was first identified as a heterochronic gene coordinating the "larva-to-adult" switch (L/A switch) in *C. elegans (1, 2, 16).* In *C. elegans,* development follows a rigid schedule of cell division and differentiation, with four specific and well-characterized larval stages (L1-L4). Lin28 expression is highest in the early embryo and L1 stage and its expression gradually decreases through L2 and is

greatly diminished by L3/L4 stages (*3*). Mutant lin-28 worms exhibit severe developmental alterations. Loss of function lin-28 mutations cause precocity, where development occurs earlier than normal. Specifically the L2 stage is omitted and L3/L4 occurs earlier than in wild-type animals (*1*). In gain of function Lin28 mutants, worm development is retarded. Early larval stage events, including proliferation of lateral hypodermal seam cells, and vulva precursor cells and other L2-specific events are reiterated (*3*). Thus, proper Lin28 expression is a prerequisite for normal development.

The Lin28 protein contains the unusual pairing of a cold-shock domain (CSD) and two zinc finger (ZnF) motifs (**Figure 1.1**). CSDs bind nucleic acids and are found in proteins in archea, bacteria and eukayotes (*17*). ZnFs are often found in proteins that bind zinc and interact by finger-like projections with DNA or RNA (*18, 19*). Lin28 is the only animal protein known to contain both a CSD and ZnFs. In plants, the *Arabidopsis thaliana* glycine-rich protein 2 (GRP2) also contains both CSD and ZnF domains. GRP2 interacts with G-rich nucleic acids and plays a prominent role in flower and seed development (*20, 21*).

Lin28 homologues have been identified in all bilaterian animals. Its expression suggests that Lin28 is an evolutionarily conserved regulator of differentiation. In Drosophila, Ciona and Xenopus, Lin28 is expressed during early embryogenesis, but is not detected in adult animals (*22*). This same pattern is observed in mammals, Lin28 is abundantly expressed in embryos and stem cells, but absent in most terminally differentiated cells with the exception of skeletal and cardiac muscle (*22, 23*).

Figure 1.1 Lin28 domain structure. Lin28 protein contains a cold-shock domain (CSD) and two zinc finger (ZnF) motifs. Human Lin28A is 209 amino acids long.

In mammals Lin28 has two functionally redundant paralogs, Lin28 (also known as Lin28A) and Lin28B. Lin28 is predominately located in the cytoplasm, although several studies that use cellular fractionation and immunofluorescence microscopy suggest some nuclear localization (*24, 25*). Lin28 binds to RNA (*22, 23, 26*). Lin28 associates with actively translating mRNAs based on polysome profiling (*23*). Based on pull down experiments, Lin28 forms complexes with mRNA binding proteins (*26*). Mutations in both the CSD and ZnFs lead to nuclear accumulation (*26*). Lin28B contains a nuclear localization sequence and may bind to primary let-7 miRNA transcripts (pri-miRNAs) in the nucleus (*24*).

In embryonic stem cells (ESCs) Lin28 contributes to pluripotency by promoting self-renewal and blocking differentiation (*11, 14, 15, 27*). Lin28 is part of a small unique set of proteins that include OCT4, SOX2, and Nanog that can reprogram terminally differentiated fibroblasts to induced pluripotent stem (iPS) cells that resemble ESCs (*28*). Lin28 is the only protein that is not a transcription factor identified thus far to assist in reprogramming. This suggests that Lin28 targeted RNAs regulate stem cell properties.

Genome-wide-association studies (GWAS) have identified single nucleotide polymorphisms (SNPs) in Lin28 and Lin28B as genetic factors regulating complex traits such as height, onset of puberty, and age of menarche in females (*29-31*). Lin28 transgenic mouse models have also shown many of these traits (*7*). These findings suggest a role for Lin28 in metabolism. In fact, in muscle cells Lin28 expression increases glucose uptake via activation of the PI3K-mTOR pathway (*6, 7*).

Lin28 overexpression is associated with many human cancers and is linked to poor prognosis (*27*). Approximately 15% of primary tumors and human cancer cell lines have increased Lin28 or Lin28B expression (*8, 9, 27, 32*). Elevated Lin28 is associated with aggressiveness of germ cell tumors, hepatocellular carcinoma, and cancers of the breast, colon, thyroid, esophagus, ovaries, prostate and head and neck (*8, 9, 33-35*). Ectopic Lin28 expression facilitates cellular transformation, and Lin28-overexpressing cells can form tumors in mice (*9*). Expression of the Lin28 target let-7 inhibits tumor formation and metastasis (*36-38*).

The Lin28-let-7 interaction

Lin28 suppresses the processing of the let-7 family of microRNAs (miRNAs) (*11, 39, 40*). miRNAs post-transcriptionally regulate gene expression by suppressing translation and/or accelerating degradation of mRNAs that bear partially complementary sequences (*41*). pri-miRNAs are stem-loop containing RNAs that are first processed by Microprocessor, a multi-enzyme complex composed of the RNAse III enzyme Drosha and its RNA binding partner DGCR8 (*42, 43*) (**Figure 1.2**). Drosha cleaves pri-miRNAs into ~70 nucleotide (nt) stem-loop-containing pre-miRNAs that are exported to the cytoplasm, where another RNAse III enzyme Dicer removes the loop generating ~22 nt imperfectly paired double-stranded RNA duplexes. One strand of the duplex is incorporated into the RNA-induced silencing complex (RISC), which binds mRNAs with partially

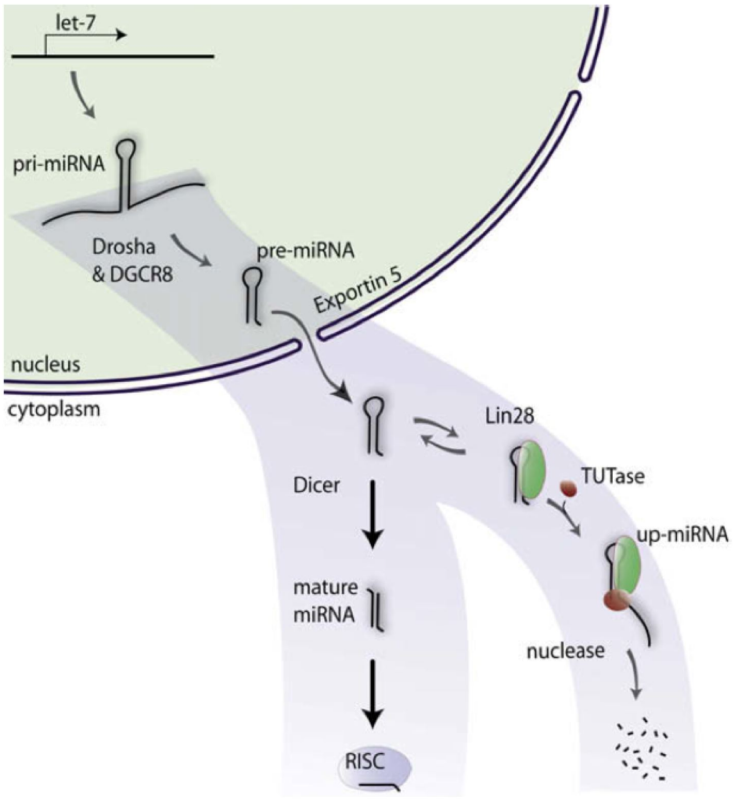


Figure 1.2 let-7 miRNA biogenesis. Primary let-7 miRNAs are transcribed in the nucleus and first processed by the RNAse III enzyme Drosha into 70 nt stem-loop-containing precursor let-7 (pre-let-7) miRNAs. Pre-let-7 miRNAs are exported to the cytoplasm and can be further processed by the RNAse III enzyme Dicer into mature let-7 miRNAs that can be loaded onto the RNA-induced silencing complex (RISC). Alternatively, Lin28 can bind the loop sequences of pre-let-7s, to recruit TUT4 which uridinylates pre-let-7s (known as up-let-7 miRNAs) to promote their degradation (*44*). Figure from Heo, I., et al (2008).

complementary sequences to induce mRNA decay and/or repress translation (*41*).

Humans express 12 let-7 miRNA family members. Each isoform contains identical nucleotides 2-8 in the mature miRNA. This region, known as the "seed sequence" (**Figure 1.3**), defines familes of miRNAs and is important in determining target mRNA specificity. miRNAs of the same family regulate overlapping, but probably not identical, sets of genes (*40*). The let-7 family functions as tumor suppressors by down regulating the expression of several oncogenes and cell cycle progression genes including c-MYC, HMGA2, RAS, cyclin D1 and CDK6 (*40, 45*).

Lin28 binds to the loops of pre-let-7 miRNAs to block let-7 processing by Dicer (*11-13*). The loop sequences are diverse in size and sequence (**Figure 1.3**) Upon binding to let-7 pre-miRNAs, Lin28 recruits ZCCHC11, also known as TUTase4 (TUT4), a uridylyl transferase, that adds a 3'uridine tail to pre-let-7s (*12, 44, 46, 47*) (**Figure 1.2**). Terminal uridylation blocks Dicer processing and promotes pre-let-7 degradation. Although most studies of Lin28 have focused on the role of Lin28 regulating let-7, the let-7 miRNAs represent less than 0.05% of the RNA that is bound to Lin28 (*48*).

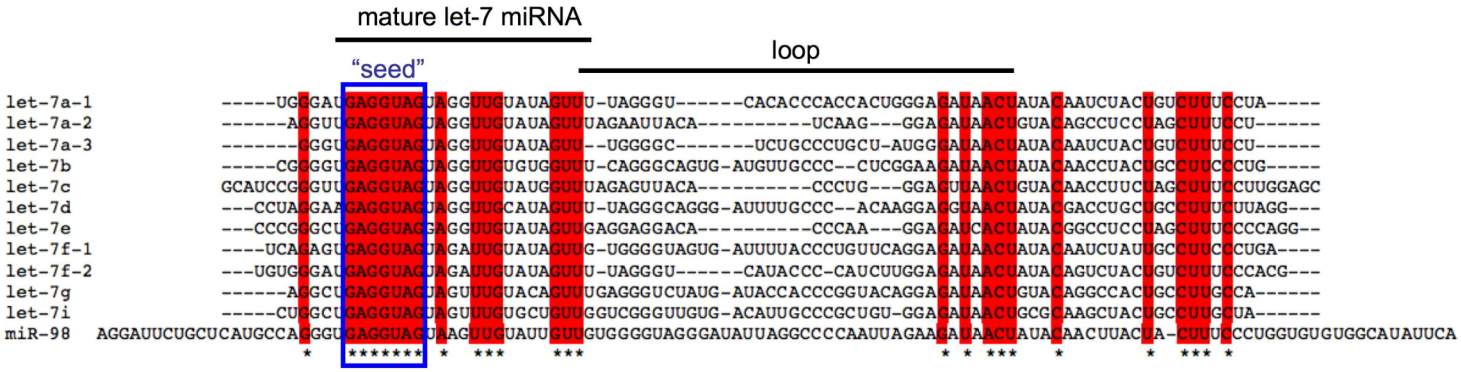


Figure 1.3 The family of pre-let-7 miRNAs. In humans precursor let-7s are processed into 10 mature let-7 miRNAs. Mature let-7 miRNAs share the same "seed sequence". The loop sequences that bind to Lin28 are variable. Conserved nucleotides are highlighted in red.

Lin28-mRNA interactions

Lin28 binds to Lin28 responsive elements (LREs) within mRNA targets either to enhance or inhibit translation. In differentiating myoblasts, Lin28 binds to the 5'untranslated region (UTR) of IGF2 mRNA to enhance IGF2 translation (*23*). In stem cells, using luciferase reporters, LREs, ranging from 95 nts to 317 nts in length, were identified in the 3'UTR and coding regions of mRNAs associated with pluripotency and cell cycle progression including OCT4, cyclin A, cyclin B, CDK4, CDK6, and H2A (*14, 15, 49*). Lin28 recruits RNA helicase A (RHA) to these mRNAs to enhance their translation (*50*). RHA is a DEAD-box RNA helicase that facilitates RNA-RNA and RNA-protein rearrangements. RHA promotes translation of a wide variety of mRNAs with highly structured 5'UTRs (*50*). Sequencing of RNAs bound to Lin28 suggested that Lin28 potentially interacts with 1,200 to 13,000 mRNAs (*5, 25, 48, 51*). How many of these mRNAs are actually regulated by Lin28 is unclear. Various methods have been used to experimentally validate Lin28 targets (**Table 1.1**). Some targets have been proposed based on protein levels that change with Lin28 knockdown and/or overexpression by western blot (*5, 48, 49, 51*). Other putative targets have been identified by qRT-PCR comparing amounts of mRNAs enriched in a Lin28 immunoprecipitation (IP) compared to pre-immune IP (*5, 49*). LREs have been verified for some mRNAs by cloning sequence elements into the 3'UTR of luciferase reporters (*5, 15, 49, 52*). Electrophoretic mobility shift analysis (EMSA) confirmed Lin28 binding to a subset of targets (*48, 52*). From these results no consistent sequence features of putative LREs has been identified.

Table 1.1 List of experimentally validated Lin28 binding RNAs (Continued). Chemical shift perturbations of let-7 resonances were monitored with NMR or heat exchange with ITC (isothermal titration calorimetry) upon Lin28 addition. EMSA (electrophoretic mobility shift analysis) confirmed direct binding of Lin28 to RNAs. IP (immunoprecipitation) was used to identify Lin28-bound RNAs in a pull-down. For many targets the Lin28 recognition mechanism has not been determined and is N/A (not available).

Lin28 binding RNA	LRE	Lin28 effect	Evidence of Lin28 binding	Proposed Lin28 recognition mechanism	Reference
let-7 miRNAs	pre-let-7 loops	Inhibits miRNA processing	EMSA, footprinting, X-ray, NMR, ITC, molecular dynamics, fluorescence	3‘GGAG	Viswanathan, S., et al (2008) Science Newman, M., et al (2008) RNA Lightfoot, H.L., et al (2011) Biochemistry Nam, Y., et al. (2011) Cell Loughlin, F.E., et al (2011) NSMB Shaik Syed Ali, P., et al (2012) FEBS Mayr, F., et al (2012) NAR Wilbert M., et al. (2012) Mol Cell Cho, J., et al (2012) Cell
CD63		Promotes translation	qRT-PCR Lin28 IP	N/A	Peng, S.,et al (2011) Stem Cells
CDK4	261 nt (3’UTR)	Promotes translation	qRT-PCR Lin28 IP, western blot, luciferase reporter	N/A	Xu, B., et al (2009) RNA Xu, B., et al (2009) NAR
Cyclin A		Promotes translation	qRT-PCR Lin28 IP, western blot	N/A	Xu, B., et al (2009) RNA Xu, B., et al (2009) NAR
Cyclin B	317 nt (1655-1971)	Promotes translation	qRT-PCR Lin28 IP, western blot, luciferase reporter	N/A	Xu, B., et al (2009) RNA Xu, B., et al (2009) NAR
DSG2	21 nt (5333-5353)	N/A	EMSA	AAGNNG, N is any nucleotide	Cho, J., et al (2012) Cell
E-Cadherin		Inhibits translation	western blot	N/A	Cho, J., et al (2012) Cell
EEF1G	330 nt (811-1140)	Promotes translation	qRT-PCR Lin28 IP, western blot, luciferase reporter	N/A	Peng, S.,et al (2011) Stem Cells
EIF4A		Promotes translation	qRT-PCR Lin28 IP	N/A	Peng, S.,et al (2011) Stem Cells

Table 1.1 List of experimentally validated Lin28 binding RNAs (Continued).

EpCAM		Inhibits translation	western blot	N/A	Cho, J., et al (2012) Cell
FUS/TLS		Promotes translation	western blot	GGAGA	Wilbert M., et al. (2012) Mol Cell
H2A	125 nt (259-393)* tandem copies	Promotes translation	qRT-PCR Lin28 IP, luciferase reporter	A bulge	Xu, B., et al (2009) NAR
HER2	200 nt (337-536)	Promotes translation	Luciferase reporter	A bulge	Lei, X-X., et al (2011) NAR
HMGA1	129 nt (5430671)	Promotes translation	qRT-PCR Lin28 IP, western blot, EMSA, luciferase reporter	A bulge	Peng, S.,et al (2011) Stem Cells Lei, X-X., et al (2011) NAR
hnRNP-F		Promotes translation	western blot	GGAGA	Wilbert M., et al. (2012) Mol Cell
HSPA5	36 nt (1019-1054)	N/A	EMSA,	AAGNG(N), N is any nucleotide	Cho, J., et al (2012) Cell
IGF2	1164 nt leader III sequence	Promotes translation	qRT-PCR Lin28 IP, polysosome profiling, luciferase reporter	N/A	Polesskaya, A., et al (2007) G&D
LAMP1		Inhibits translation	western blot	N/A	Cho, J., et al (2012) Cell
Lin28	3'UTR	Promotes translation	western blot, luciferase reporter	GGAGA	Wilbert M., et al. (2012) Mol Cell
OCT4	95 nt (516-610)	Promotes translation	qRT-PCR Lin28 IP, polysome profiling, luciferase reporter, western blot	A bulge	Qiu, C., et al (2009) NAR Peng, S.,et al (2011) Stem Cells Lei, X-X., et al (2011) NAR
RPS13	456 nt (33-488)	Promotes translation	qRT-PCR Lin28 IP, western blot, luciferase reporter	N/A	Peng, S.,et al (2011) Stem Cells
RPS19	106 nt (486-591)	Promotes translation	Luciferase reporter	A bulge	Lei, X-X., et al (2011) NAR
SUN1	20 nt (1841-1860)	N/A	EMSA	(N)UGUG(N), N is any nucleotide	Cho, J., et al (2012) Cell
TDP-43	85 nt (2138-2222)	Promotes translation	western blot, luciferase reporter	GGAGA	Wilbert M., et al. (2012) Mol Cell
TIA-1		Promotes translation	western blot	GGAGA	Wilbert M., et al. (2012) Mol Cell

The Lin28 recognition mechanism

Mutagenesis together with *in vitro* binding assays has suggested that guanine (*13, 53, 54*), adenine (*13*), uracil (*25*), and cytosine (*55*) residues affect Lin28 binding. High-resolution crystal structures of Lin28 bound to pre-let-7 loop oligonucleotides suggested the CSD and ZnF domains form critical contacts with guanine and adenine bases in distinct G-rich regions in the loops of pre-let-7s (**Figure 1.4**) (*54, 56, 57*). In particular the ZnF domain was proposed to bind a 3'GGAG motif in the loop sequence of pre-let-7 miRNAs. However, the NMR structure of isolated ZnFs complexed with AGGAGAU suggested that an exact GGAG sequence is not required. The authors proposed that the ZnFs bind a consensus sequence of NGNNG, where N is any nucleotide (*56*). Binding and structural studies of the isolated Lin28 CSD suggested that the CSD has very little sequence specificity since it formed crystals with $d(T)_6$ and $r(U)_6$ (*57*). Sequencing of Lin28-bound mRNAs has led to suggestions that putative Lin28 consensus sequences might include GGAGA, AAGNNG, AAGNG(N) and (N)UGUG(N) where N is any nucleotide (*48, 51*). A molecular modeling study instead suggested that Lin28 recognizes a critical "A" bulge flanked by two G:C bonds in an extended double stranded region of the LREs of mRNAs (**Figure 1.5**) (*52*). A single A to U mutation in mRNAs of OCT4, HMGA1, and RPS19 inhibited Lin28's ability to stimulate their translation (*52*). Since the "A bulge" is not present in pre-let-7 miRNAs, these authors suggested that Lin28 might recognize distinct features in its miRNA and mRNA targets.

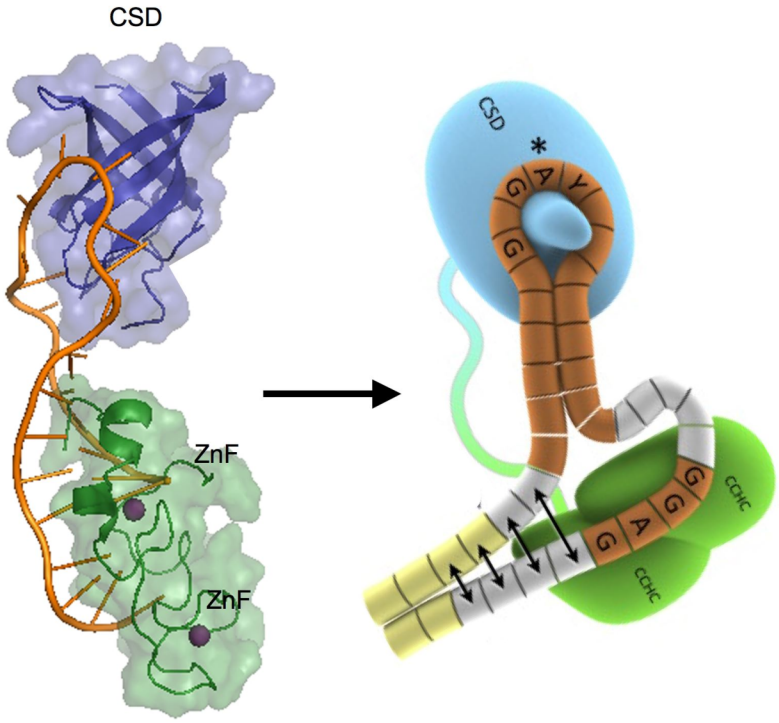


Figure 1.4 Lin28 binding to pre-let-7 elements. The crystal structure of mouse Lin28A bound to elements of the pre-let-7 loop sequence (shown in orange) suggested that the CSD and ZnF fingers bind critical guanine and adenine nucleotides. The CSD was proposed to bind a GNGAY sequence where N is any nucleotide and Y is a pyrimidine. The CCHC ZnFs were suggested to bind a 3'GGAG motif. A three-dimensional model was generated based on the Lin28 structure bound to the loop of pre-let-7g (PDB ID 3TS2). The schematic for Lin28 binding is adapted from Nam, Y., et al. (2011). The CSD is in blue, the ZnFs are in green, the pre-let-7 loop oligonucleotides are in orange and nucleotides extending into the pre-let-7 stem are indicated in yellow.

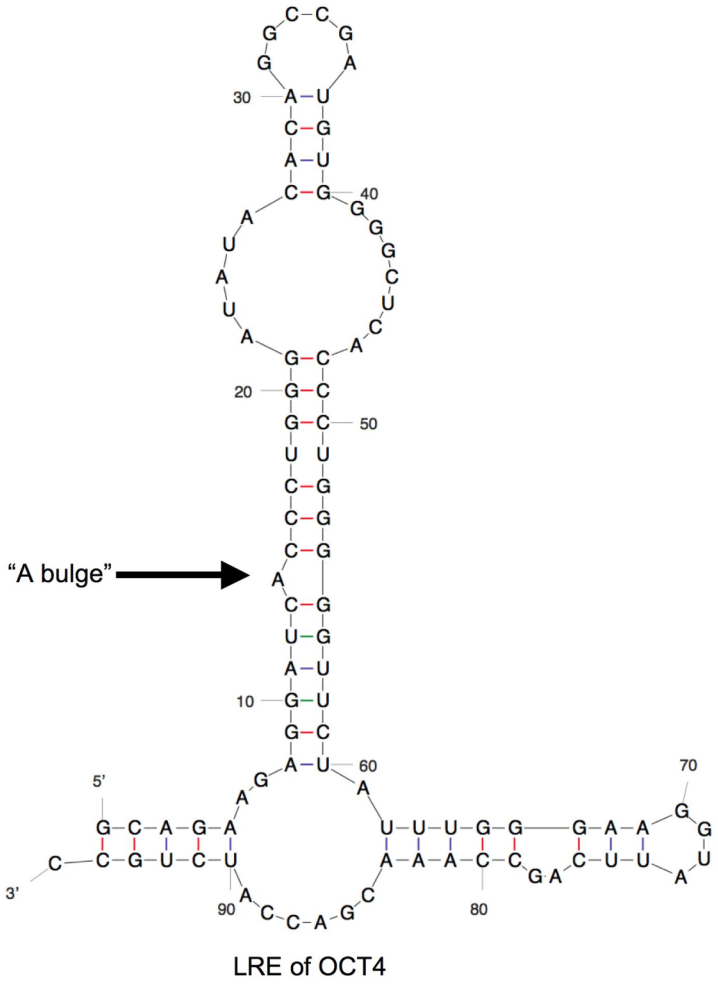


Figure 1.5 A bulge of Lin28 mRNA targets. Lei, X-X., et al (2011) proposed that Lin28 binds to a critical unpaired adenine ("A bulge") flanked by G:C bonds in an extended double stranded region in the Lin28 responsive elements (LREs) of mRNA targets. An example is the LRE of OCT4.

These structural, sequencing and modeling data have suggested short Lin28 recognition elements. Such sequences are abundant in mRNAs throughout the transcriptome and hence are unlikely to uniquely define Lin28-binding specificity. None of these motifs are useful for predicting LREs of Lin28-regulated mRNAs. Thus, the key RNA features that determine Lin28 binding remain unclear.

Thesis Objectives:

The central hypothesis behind this work is that Lin28 regulated miRNAs and mRNAs share common structural features. To address this, we analyzed the structural features of the pre-let-7s and their loops which contain the 3'GGAG motif; the LREs of DSG2, HSPA5 and SUN1, which contain the putative Lin28 consensus sequences AAGNNG, AAGNG(N), and (N)UGUG(N), respectively; and the LRE of OCT4 which contains the proposed "A bulge".

We found that pre-let-7 loops and Lin28 recognition elements (LREs) have an abundance of guanine and guanine repeats. Because guanine rich nucleic acids can form stable planar structures called G-quartets by G:G:G:G hydrogen bonding, we hypothesized that Lin28 might recognize structured RNAs that contain G-quartets. To examine this hypothesis we used NMR spectroscopy, gel shift assays and a dye that specifically binds to G-quartets. We provide evidence that a G-quartet structure is important for Lin28 recognition and might be a unifying feature of Lin28 RNA targets.

References:

1. Ambros, V., and Horvitz, H. R. (1984) Heterochronic mutants of the nematode Caenorhabditis elegans, *Science 226*, 409-416.

2. Liu, Z. C., and Ambros, V. (1989) Heterochronic genes control the stage-specific initiation and expression of the dauer larva developmental program in Caenorhabditis elegans, *Genes Dev 3*, 2039-2049.

3. Moss, E. G., Lee, R. C., and Ambros, V. (1997) The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA, *Cell 88*, 637-646.

4. Zheng, K., Wu, X., Kaestner, K. H., and Wang, P. J. (2009) The pluripotency factor LIN28 marks undifferentiated spermatogonia in mouse, *BMC Developmental Biology 9*, 38-49.

5. Peng, S., Chen, L.-L., Lei, X.-X., Yang, L., Lin, H., Carmichael, G. G., and Huang, Y. (2011) Genome-Wide Studies Reveal That Lin28 Enhances the Translation of Genes Important for Growth and Survival of Human Embryonic Stem Cells, *Stem Cells 29*, 496-504.

6. Zhu, H., Shyh-Chang, N., Segrè, A. V., Shinoda, G., Shah, S. P., Einhorn, W. S., Takeuchi, A., Engreitz, J. M., Hagan, J. P., Kharas, M. G., Urbach, A., Thornton, J. E., Triboulet, R., Gregory, R. I., Consortium, D., Investigators, M., Altshuler, D., and Daley, G. Q. (2011) The Lin28/let-7 axis regulates glucose metabolism, *Cell 147*, 81-94.

7. Zhu, H., Shah, S., Shyh-Chang, N., Shinoda, G., Einhorn, W. S., Viswanathan, S. R., Takeuchi, A., Grasemann, C., Rinn, J. L., Lopez, M. F., Hirschhorn, J. N., Palmert, M. R., and Daley, G. Q. (2010) Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies, *Nat Genet 42*, 626-630.

8. West, J. A., Viswanathan, S. R., Yabuuchi, A., Cunniff, K., Takeuchi, A., Park, I.-H., Sero, J. E., Zhu, H., Perez-Atayde, A., Frazier, A. L., Surani, M. A., and Daley, G. Q. (2009) A role for Lin28 in primordial germ-cell development and germ-cell malignancy, *Nature 460*, 909-913.

9. Viswanathan, S. R., Powers, J. T., Einhorn, W., Hoshida, Y., Ng, T. L., Toffanin, S., O'Sullivan, M., Lu, J., Phillips, L. A., Lockhart, V. L., Shah, S. P., Tanwar, P. S., Mermel, C. H., Beroukhim, R., Azam, M., Teixeira, J., Meyerson, M., Hughes, T. P., Llovet, J. M., Radich, J., Mullighan, C. G., Golub, T. R., Sorensen, P. H., and Daley, G. Q. (2009) Lin28 promotes transformation and is associated with advanced human malignancies, *Nat Genet 41*, 843-848.

10. Meirelles, K., Benedict, L. A., Dombkowski, D., Pepin, D., Preffer, F. I., Teixeira, J., Tanwar, P. S., Young, R. H., MacLaughlin, D. T., Donahoe, P. K., and Wei, X. (2012) Human ovarian cancer stem/progenitor cells are stimulated by doxorubicin but inhibited by Mullerian inhibiting substance, *Proc Natl Acad Sci USA 109*, 2358-2363.

11. Viswanathan, S. R., Daley, G. Q., and Gregory, R. I. (2008) Selective blockade of microRNA processing by Lin28, *Science 320*, 97-100.

12. Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V. N. (2008) Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA, *Mol Cell 32*, 276-284.

13. Newman, M. A., Thomson, J. M., and Hammond, S. M. (2008) Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing, *RNA 14*, 1539-1549.

14. Qiu, C., Ma, Y., Wang, J., Peng, S., and Huang, Y. (2010) Lin28-mediated post-transcriptional regulation of Oct4 expression in human embryonic stem cells, *Nucleic Acids Res 38*, 1240-1248.

15. Xu, B., Zhang, K., and Huang, Y. (2009) Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells, *RNA 15*, 357-361.

16. Ambros, V. (1989) A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans, *Cell 57*, 49-57.

17. Mihailovich, M., Militti, C., Gabaldón, T., and Gebauer, F. (2010) Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression, *Bioessays 32*, 109-118.

18. Razin, S. V., Borunova, V. V., Maksimenko, O. G., and Kantidze, O. L. (2012) Cys2His2 zinc finger protein family: classification, functions, and major members, *Biochemistry Mosc 77*, 217-226.

19. Muriaux, D., and Darlix, J.-L. (2010) Properties and functions of the nucleocapsid protein in virus assembly, *RNA Biol 7*, 744-753.

20. Kim, J. S., Park, S. J., Kwak, K. J., Kim, Y. O., Kim, J. Y., Song, J., Jang, B., Jung, C.-H., and Kang, H. (2007) Cold shock domain proteins and glycine-rich RNA-binding proteins from Arabidopsis thaliana can promote the cold adaptation process in Escherichia coli, *Nucleic Acids Res 35*, 506-516.

21. Kingsley, P. D., and Palis, J. (1994) GRP2 proteins contain both CCHC zinc fingers and a cold shock domain, *Plant Cell 6*, 1522-1523.

22. Moss, E. G., and Tang, L. (2003) Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites, *Dev Biol 258*, 432-442.

23. Polesskaya, A., Cuvellier, S., Naguibneva, I., Duquet, A., Moss, E. G., and Harel-Bellan, A. (2007) Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency, *Genes Dev 21*, 1125-1138.

24. Piskounova, E., Polytarchou, C., Thornton, J. E., LaPierre, R. J., Pothoulakis, C., Hagan, J. P., Iliopoulos, D., and Gregory, R. I. (2011) Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms, *Cell 147*, 1066-1079.

25. Hafner, M., Max, K. E. A., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013) Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition, *RNA 19*, 1-14.

26. Balzer, E., and Moss, E. G. (2007) Localization of the developmental timing regulator Lin28 to mRNP complexes, P-bodies and stress granules, *RNA Biol 4*, 16-25.

27. Viswanathan, S. R., and Daley, G. Q. (2010) Lin28: A microRNA regulator with a macro role, *Cell 140*, 445-449.

28. Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2007) Induced pluripotent stem cell lines derived from human somatic cells, *Science 318*, 1917-1920.

29. Ong, K. K., Elks, C. E., Li, S., Zhao, J. H., Luan, J. a., Andersen, L. B., Bingham, S. A., Brage, S., Smith, G. D., Ekelund, U., Gillson, C. J., Glaser, B., Golding, J., Hardy, R., Khaw, K.-T., Kuh, D., Luben, R., Marcus, M., McGeehin, M. A., Ness, A. R., Northstone, K., Ring, S. M., Rubin, C., Sims, M. A., Song, K., Strachan, D. P., Vollenweider, P., Waeber, G., Waterworth, D. M., Wong, A., Deloukas, P., Barroso, I., Mooser, V., Loos, R. J., and Wareham, N. J. (2009) Genetic variation in LIN28B is associated with the timing of puberty, *Nat Genet 41*, 729-733.

30. Perry, J. R. B., Stolk, L., Franceschini, N., Lunetta, K. L., Zhai, G., McArdle, P. F., Smith, A. V., Aspelund, T., Bandinelli, S., Boerwinkle, E., Cherkas, L., Eiriksdottir, G., Estrada, K., Ferrucci, L., Folsom, A. R., Garcia, M., Gudnason, V., Hofman, A., Karasik, D., Kiel, D. P., Launer, L. J., van Meurs, J., Nalls, M. A., Rivadeneira, F., Shuldiner, A. R., Singleton, A., Soranzo, N., Tanaka, T., Visser, J. A., Weedon, M. N., Wilson, S. G., Zhuang, V., Streeten, E. A., Harris, T. B., Murray, A., Spector, T. D., Demerath, E. W., Uitterlinden, A. G., and Murabito, J. M. (2009) Meta-

analysis of genome-wide association data identifies two loci influencing age at menarche, *Nat Genet 41*, 648-650.

31. Sulem, P., Gudbjartsson, D. F., Rafnar, T., Holm, H., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Alexandersen, P., Feenstra, B., Boyd, H. A., Aben, K. K., Verbeek, A. L. M., Roeleveld, N., Jonasdottir, A., Styrkarsdottir, U., Steinthorsdottir, V., Karason, A., Stacey, S. N., Gudmundsson, J., Jakobsdottir, M., Thorleifsson, G., Hardarson, G., Gulcher, J., Kong, A., Kiemeney, L. A., Melbye, M., Christiansen, C., Tryggvadottir, L., Thorsteinsdottir, U., and Stefansson, K. (2009) Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche, *Nat Genet 41*, 734-738.

32. Iliopoulos, D., Hirsch, H. A., and Struhl, K. (2009) An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation, *Cell 139*, 693-706.

33. King, C. E., Cuatrecasas, M., Castells, A., Sepulveda, A. R., Lee, J.-S., and Rustgi, A. K. (2011) LIN28B promotes colon cancer progression and metastasis, *Cancer Res 71*, 4260-4268.

34. Xue, D., Peng, Y., Wang, F., Allan, R. W., and Cao, D. (2011) RNA-binding protein LIN28 is a sensitive marker of ovarian primitive germ cell tumours, *Histopath 59*, 452-459.

35. Dangi-Garimella, S., Yun, J., Eves, E. M., Newman, M., Erkeland, S. J., Hammond, S. M., Minn, A. J., and Rosner, M. R. (2009) Raf kinase inhibitory protein suppresses a metastasis signalling cascade involving LIN28 and let-7, *EMBO J 28*, 347-358.

36. Yu, F., Yao, H., Zhu, P., Zhang, X., Pan, Q., Gong, C., Huang, Y., Hu, X., Su, F., Lieberman, J., and Song, E. (2007) let-7 regulates self renewal and tumorigenicity of breast cancer cells, *Cell 131*, 1109-1123.

37. Chang, T.-C., Zeitels, L. R., Hwang, H.-W., Chivukula, R. R., Wentzel, E. A., Dews, M., Jung, J., Gao, P., Dang, C. V., Beer, M. A., Thomas-Tikhonenko, A., and Mendell, J. T. (2009) Lin-28B transactivation is necessary for Myc-mediated let-7 repression and proliferation, *Proc Natl Acad Sci USA 106*, 3384-3389.

38. Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D., and Slack, F. J. (2005) RAS is regulated by the let-7 microRNA family, *Cell 120*, 635-647.

39. Abbott, A. L., Alvarez-Saavedra, E., Miska, E. A., Lau, N. C., Bartel, D. P., Horvitz, H. R., and Ambros, V. (2005) The let-7 MicroRNA family members

mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans, *Dev Cell 9*, 403-414.

40. Büssing, I., Slack, F. J., and Grosshans, H. (2008) let-7 microRNAs in development, stem cells and cancer, *Trends Mol Med 14*, 400-409.

41. Dugas, D. V., and Bartel, B. (2004) MicroRNA regulation of gene expression in plants, *Curr Opin Plant Biol 7*, 512-520.

42. Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., and Kim, V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II, *EMBO J 23*, 4051-4060.

43. Gregory, R. I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004) The Microprocessor complex mediates the genesis of microRNAs, *Nature 432*, 235-240.

44. Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., and Kim, V. N. (2009) TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation, *Cell 138*, 696-708.

45. Johnson, C. D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., Wilson, M., Wang, X., Shelton, J., Shingara, J., Chin, L., Brown, D., and Slack, F. J. (2007) The let-7 microRNA represses cell proliferation pathways in human cells, *Cancer Res 67*, 7713-7722.

46. Hagan, J. P., Piskounova, E., and Gregory, R. I. (2009) Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells, *Nat Struct Mol Biol 16*, 1021-1025.

47. Lehrbach, N. J., Armisen, J., Lightfoot, H. L., Murfitt, K. J., Bugaut, A., Balasubramanian, S., and Miska, E. A. (2009) LIN-28 and the poly(U) polymerase PUP-2 regulate let-7 microRNA processing in Caenorhabditis elegans, *Nat Struct Mol Biol 16*, 1016-1020.

48. Cho, J., Chang, H., Kwon, S. C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y. K., and Kim, V. N. (2012) LIN28A Is a Suppressor of ER-Associated Translation in Embryonic Stem Cells, *Cell 151*, 765-777.

49. Xu, B., and Huang, Y. (2009) Histone H2a mRNA interacts with Lin28 and contains a Lin28-dependent posttranscriptional regulatory element, *Nucleic Acids Res 37*, 4256-4263.

50. Jin, J., Jing, W., Lei, X.-X., Feng, C., Peng, S., Boris-Lawrie, K., and Huang, Y. (2011) Evidence that Lin28 stimulates translation by recruiting RNA helicase A to polysomes, *Nucleic Acids Res 39*, 3724-3734.

51. Wilbert, M. L., Huelga, S. C., Kapeli, K., Stark, T. J., Liang, T. Y., Chen, S. X., Yan, B. Y., Nathanson, J. L., Hutt, K. R., Lovci, M. T., Kazan, H., Vu, A. Q., Massirer, K. B., Morris, Q., Hoon, S., and Yeo, G. W. (2012) LIN28 Binds Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance, *Mol Cell 48*, 1-12.

52. Lei, X.-X., Xu, J., Ma, W., Qiao, C., Newman, M. A., Hammond, S. M., and Huang, Y. (2011) Determinants of mRNA recognition and translation regulation by Lin28, *Nucleic Acids Res 40*, 3574-3584.

53. Lightfoot, H. L., Bugaut, A., Armisen, J., Lehrbach, N. J., Miska, E. A., and Balasubramanian, S. (2011) A LIN28-dependent structural change of pre-let-7g directly inhibits Dicer processing, *Biochemistry 50*, 7514-7521.

54. Nam, Y., Chen, C., Gregory, Richard I., Chou, James J., and Sliz, P. (2011) Molecular Basis for Interaction of let-7 MicroRNAs with Lin28, *Cell 147*, 1080-1091.

55. Piskounova, E., Viswanathan, S. R., Janas, M., LaPierre, R. J., Daley, G. Q., Sliz, P., and Gregory, R. I. (2008) Determinants of microRNA processing inhibition by the developmentally regulated RNA-binding protein Lin28, *J Biol Chem 283*, 21310-21314.

56. Loughlin, F. E., Gebert, L. F. R., Towbin, H., Brunschweiger, A., Hall, J., and Allain, F. H.-T. (2011) Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28, *Nature Struct Mol Bio 19*, 84-89.

57. Mayr, F., Schütz, A., Döge, N., and Heinemann, U. (2012) The Lin28 cold-shock domain remodels pre-let-7 microRNA, *Nucleic Acids Res*, 1-15.

Chapter 2: Lin28 regulated miRNAs and mRNAs contain G-quartet features

let-7 loops and mRNAs with LREs bind to Lin28 by electrophorectic mobility shift analysis

We first confirmed that the pre-let-7s (pre-let-7a and pre-let-7g), the let-7 loops (let-7g loop and let-7d loop) and the LREs of DSG2, HSPA5, SUN1 and OCT4 bind Lin28 by electrophoretic mobility shift analysis (EMSA). ^{32}P-labeled target RNAs were incubated with increasing amounts of recombinant mouse Lin28A (rLin28A) (**Figure 2.1**). Gel shifts increased as more rLin28A was added to the reaction. Thus, Lin28A directly binds to each of these putative Lin28 targets. A ^{32}P-labeled miR-21 loop was used as a negative control and even at the highest Lin28 concentration did not form a gel shift.

1D NMR spectra of Lin28 binding RNAs contain resonances suggestive of non-canonical hydrogen bonds

We utilized NMR spectroscopy to gain insight into the structural features of Lin28 target RNAs. Imino protons are the protons attached to nitrogen adjacent to a carbonyl. By monitoring imino protons in nucleic acids, NMR can provide information about the hydrogen-bonding network that governs nucleic acid structure. Resonances from labile imino protons are only observable when they are protected from solvent exchange by hydrogen bonding (**Figure 2.2**). In Watson-Crick base pairing, the imino proton of uracil in an A:U hydrogen bond has a chemical shift at ~13 ppm, while the imino proton of guanine in a G:C hydrogen bond is closer to 12.7 ppm (*1*). G:U wobbles and other non-canonical hydrogen bonds occur upfield of 12.7 ppm (*1-5*). For standard A-form RNA or B-

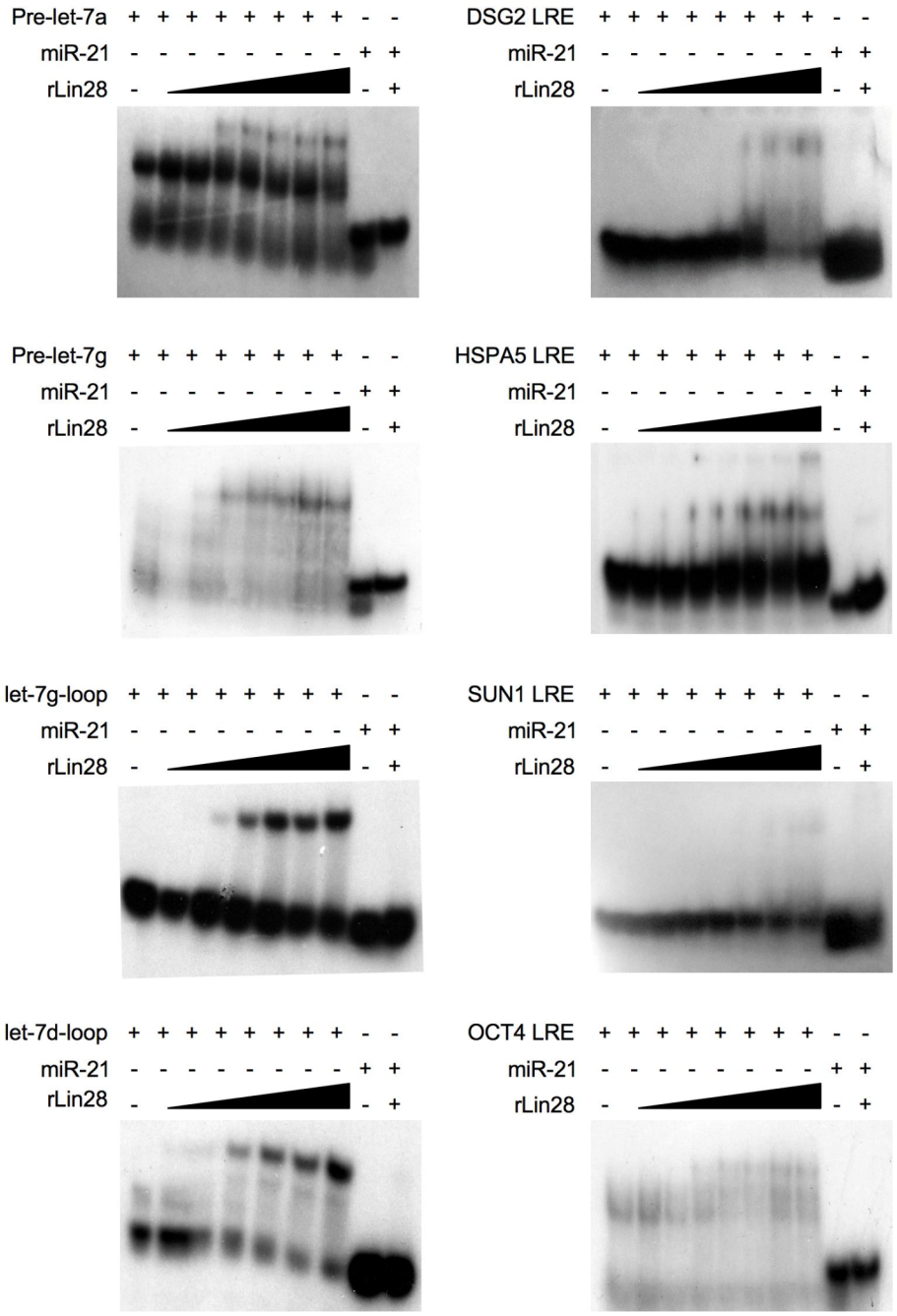


Figure 2.1 Lin28 directly binds RNA targets. Electrophoretic mobility shift analysis (EMSA) confirms the interaction between Lin28 and the target sequences (pre-let-7a, pre-let-7g, let-7g-loop, let-7d-loop, and LREs of DSG2, HSPA5, SUN1 and OCT4). Lin28-RNA complexes are visualized by autoradiography as a more slowly migrating band. ^{32}P-miR-21 did not form a complex at the highest Lin28 concentration (10 μM).

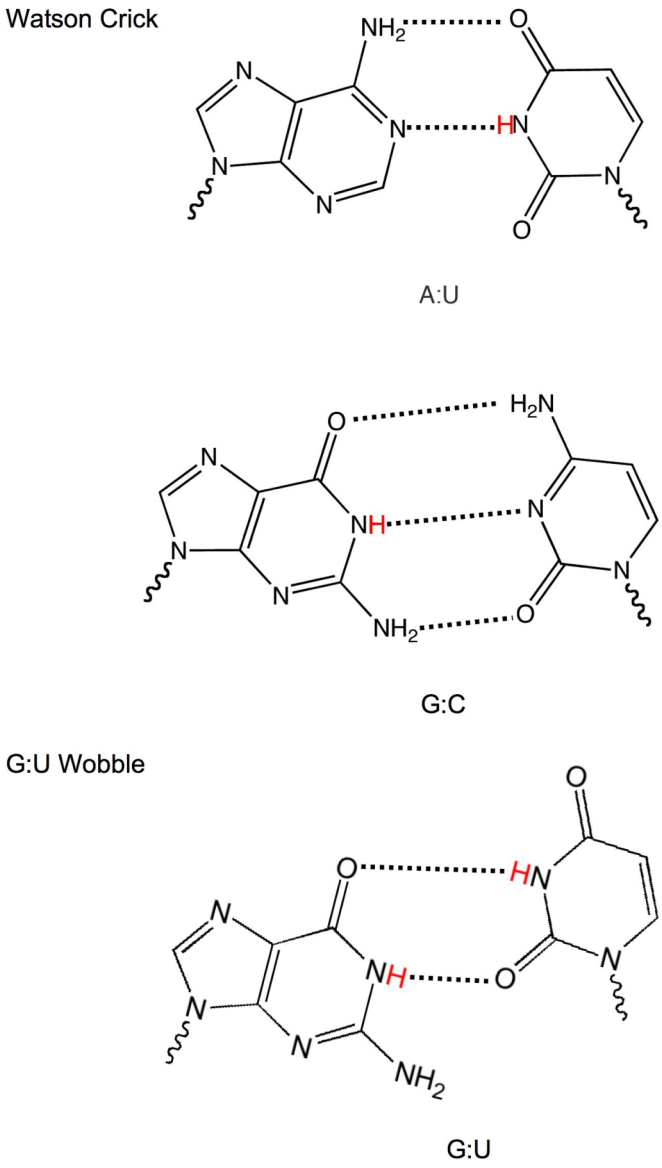


Figure 2.2 Watson-Crick and G:U wobble hydrogen bonds. Hydrogen-bonded imino protons of uracil and guanine (in red) are protected from solvent exchange and are observable by NMR spectroscopy.

form DNA, the number of imino resonances correlates with the number of A:U or G:C hydrogen bonds.

Utilizing mFOLD (*6*), a bioinformatics tool that uses a nearest neighbor approach to calculate energy minimized nucleic acid structures, we generated predicted secondary structures for the let-7a loop, the let-7g loop, and the LREs of DSG2 and OCT4 (**Figure 2.3**). All these Lin28 targets were predicted to form "stem-loop" structures with predominately Watson-Crick base pairing. To probe their actual hydrogen-bonding network, we recorded 1D proton NMR spectra for each of these Lin28-binding RNAs (**Figure 2.4**)**.** All of the RNAs had imino resonances indicating the presence of hydrogen bonds. However in addition to putative A:U and G:C bonds, each Lin28 RNA target had resonances in the region from ~10.0-12.4 ppm indicative of the formation of non-canonical hydrogen bonds.

The observed resonances in this region exceeded the number of predicted G:U wobbles based on mFOLD structures (**Figure 2.3**). This suggests that additional unusual hydrogen bonding networks might be present in these Lin28 RNA targets. Further, both let-7 loop spectra contain resonance peaks around ~9.3-9.5 ppm. Resonances in this region have been previously observed when guanine or adenine amino protons are shifted downfield due to unusual hydrogen bonding found in G:A or A:A mismatches (*7*) (**Figure 2.5**). This suggests the structure of the let-7 loops contains atypical hydrogen bonds that may include G:A or A:A binding.

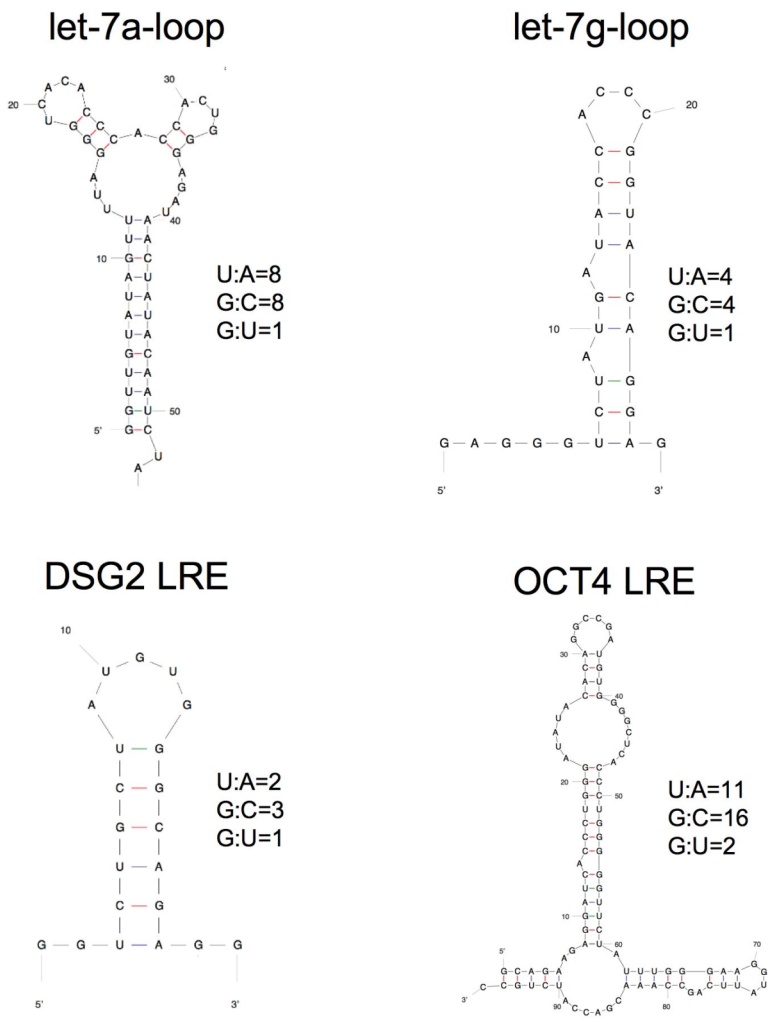


Figure 2.3 Predicted secondary structures of Lin28 targets contain diverse stem-loop structures. mFOLD-generated secondary structures for selected Lin28 targets including the let-7a loop, let-7g loop and the LREs for DSG2 and OCT4. The length of the stem and the size, number and sequence of the loops vary for each target. The number of predicted base-pairs is listed.

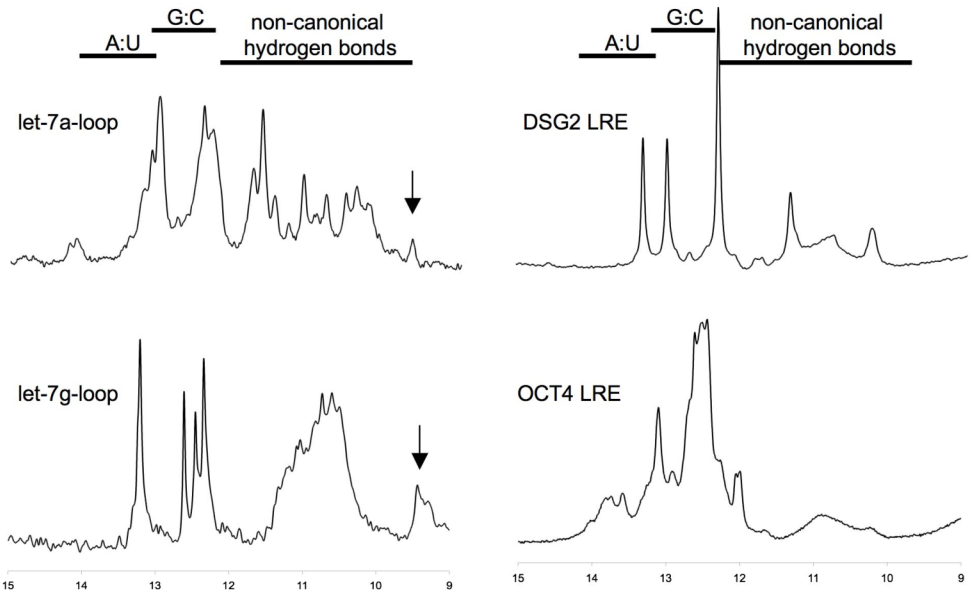


Figure 2.4 Lin28 targets contain non-canonical hydrogen bonds. The 1D imino proton spectra for Lin28 targets (the let-7a-loop, let-7g-loop and the LREs for DSG2 and OCT4) contain imino resonances indicative of non-canonical hydrogen bonds. The let-7 loop spectra contain unusual peaks at 9.3-9.5 ppm (indicated by arrows). Spectra were recorded at 5ºC.

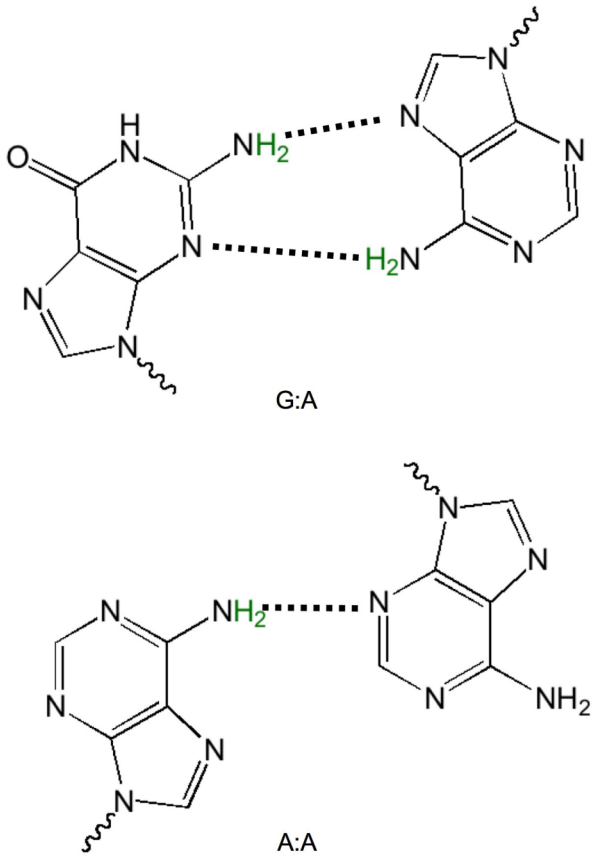


Figure 2.5 G:A and A:A hydrogen bonds. The amino proton resonances (green) of G:A and A:A hydrogen bonds have characteristic chemical shifts of ~9.3- 9.5 ppm (7).

The LRE of DSG2 is 21 nucleotides long and 48% of the LRE sequence is guanine. The DSG2 LRE spectrum contained several broad intense resonances from ~10.2-11.4 ppm and an intense peak at 12.3 ppm (**Figure 2.4**). When DSG2 LRE was concentrated above 500 μM, it formed a gel. Forming gelatinous structures at high concentrations has been observed previously for G-rich nucleic acids that contain G-quartet structures (*8*). The 95 nt LRE of OCT4 is longer than the other Lin28 LREs we studied. It is also enriched for guanines (30.5%) and contains stretches of 2-5 consecutive guanines. Due to its large size and slow tumbling rate, the spectra of OCT4 contains several poorly resolved resonances, nonetheless the broad peaks at 10.2-10.8 ppm, 11.6 ppm, and 12.0 ppm suggest that the OCT4 LRE also contains non-Watson-Crick base pairs (**Figure 2.4)**.

Because of the imino resonance pattern and high G-content, we hypothesized that these Lin28-binding RNAs might form G-quartets (G4s). G4s are composed of four guanines interconnected in a planar configuration through cyclic Hoogsteen hydrogen bonding (**Figure 2.6**). G-rich nucleic acids can form stacked G4s that adopt unique structures known as G-quadruplexes (*9-11*). The unusual hydrogen bonding pattern of G4s gives rise to characteristic imino resonances at ~10-12.2 ppm (*1, 2*). Peaks within this region are highly suggestive of G-quartet formation. From our 1D NMR analysis we observed that the let-7a loop, let-7g loop, and LREs of DSG2 and OCT4 contain imino resonances in this region, suggesting that these Lin28-binding RNAs might contain G4 features.

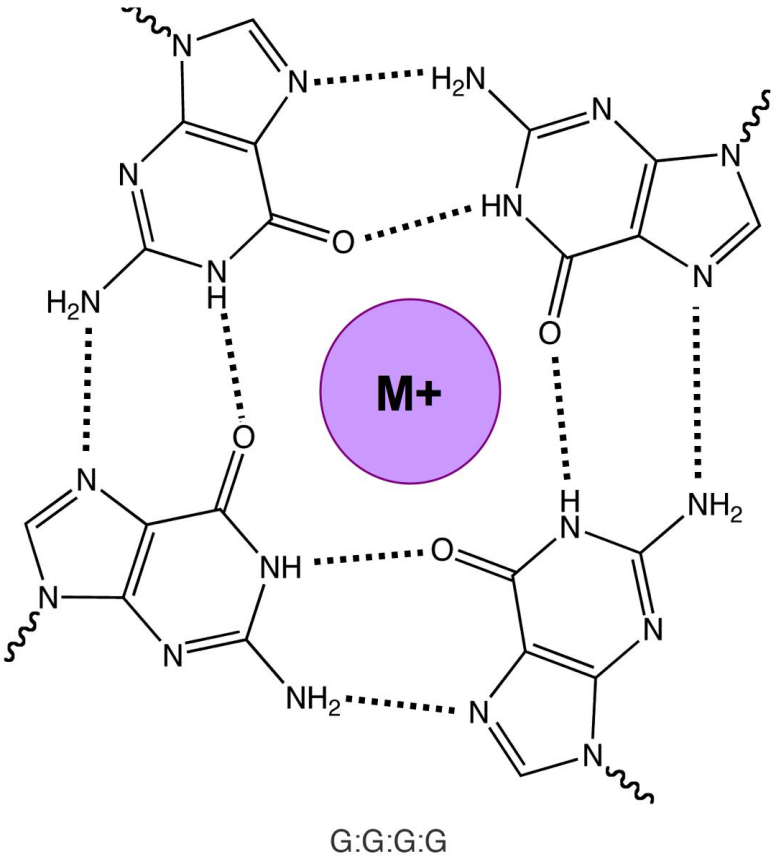


Figure 2.6 Hydrogen bonding of G-quartets. G-quartets are composed of at least four guanines interconnected through Hoogsteen hydrogen bonds. A monovalent ion (M+) binds within the central cavity of G-quartets. G-quadruplexes contain multiple stacked quartets.

NMR resonances between 10.2-11.4 ppm in the let-7g loop are thermodynamically stable.

The hydrogen bond network of G4s has increased thermostability and slower unfolding rates than standard Watson-Crick base pairs (*12*). We therefore examined whether the imino frequency resonances corresponding to the putative G4-like hydrogen bonds in the let-7g loop were stable as the temperature increased (**Figure 2.7**). By raising the temperature from 10°C-40°C the resonances attributed to A:U and G:C pairs disappeared. However the proposed G4 resonances at 10.2-12.4 ppm remained intact. This enhanced thermostability provides supporting evidence for the formation of a G4 by the let-7g loop.

Lin28 binding RNAs form higher order oligomers that prevent high-resolution NMR spectra

We attempted to use additional multidimensional NMR methods, such as Nuclear Overhauser Effect spectroscopy (NOESY), to unambiguously map the hydrogen bond connectivity of Lin28-binding RNAs. The cross peaks in a NOESY spectrum connect resonances from nuclei that are located within close proximity (~5 Å) (*13*). However despite adjustments to buffer, concentration, temperature, counter ions and pH, the resonances in NOESY spectra for Lin28-binding RNAs remained poorly defined, making the spectra difficult to interrupt. A representative NOESY spectrum from the let-7g-loop is shown in **Figure 2.8**.

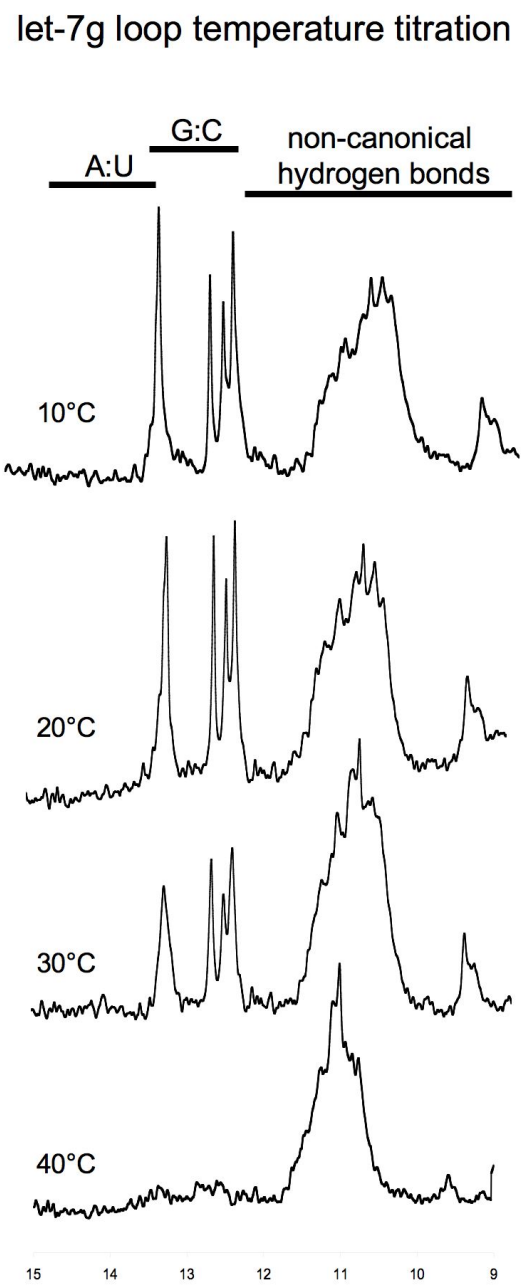


Figure 2.7 The G-quartet-like bonds of let-7g loop are thermodynamically stable. 1D proton spectra of the let-7g-loop at 10°C, 20°C, 30°C and 40°C indicate the proposed G-quartet hydrogen bonds have increased stability compared to A:U and G:C bonds whose resonances disappear at 40°C.

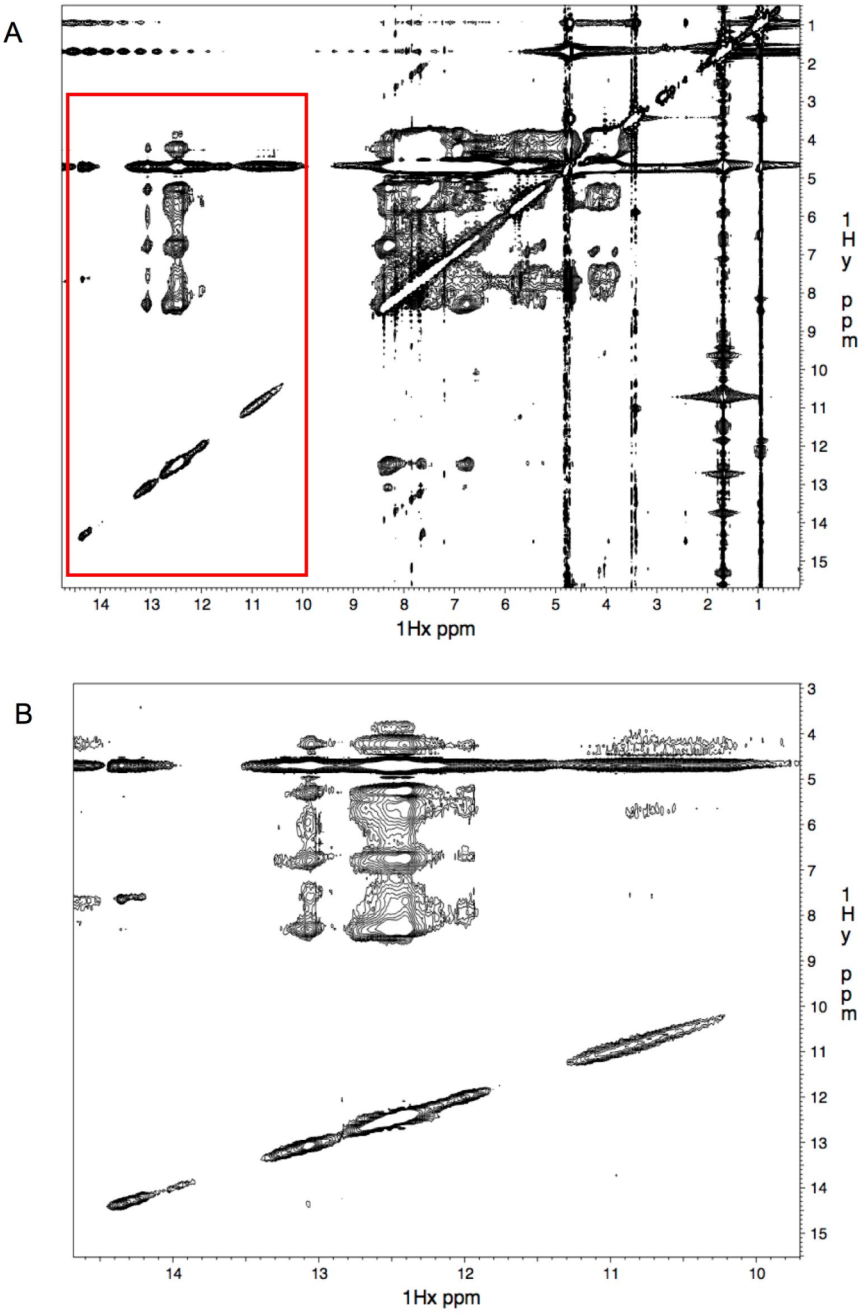


Figure 2.8 ^{1}H-^{1}H NOESY of let-7g loop. The NOESY spectra for many Lin28 targets contain weak NOES with very poor resolution. Alterations in buffer, concentration, pH, temperature and counter-ion concentrations did not significantly improve spectra. Shown is a representative ^{1}H-^{1}H NOESY spectrum of let-7g-loop (A) and close up of the imino region (B).

Because of the poor spectral resolution, Lin28-binding RNAs may oligomerize and/or dynamically change structure in solution. To examine this possibility, we analyzed the Lin28-binding RNAs by denaturing and native polyacrylamide gel electrophoresis (PAGE) (**Figure 2.9**). The Lin28 targets migrated as a single species on a denaturing gel when heated and denatured by adding 8 M Urea. However, when the same RNAs were electrophoresed under non-denaturing conditions, they formed multiple bands suggestive of higher order oligomers. Multiple bands persisted when the RNA concentration ranged from 0.05-10 μM. RNA multimers can also be sensitive to composition and concentration. Addition of Na^+, K^+, Li^+ or Mg^{2+} chloride salts did not affect the migration pattern. Thus, Lin28 targets may form multiple species. Since NMR requires high concentrations of a single RNA species to obtain quality spectra, the formation of multiple species in solution likely was responsible for the poor quality NOESY spectra.

Lin28-binding RNAs bind the G4-specific dye N-methyl mesoporphyrin IX

N-methyl mesoporphyrin IX (NMM) is a dye that binds selectively to G4s (*14-17*) (**Figure 2.10**). Based on the structure of NMM bound to human telomeric DNA, the large planar surface of the porphyrin ring of NMM interacts in parallel with planar G-quartets (*14, 15*). When it binds to G4s, NMM is excited at 399 nm and fluoresces at 614 nm (*17*). We therefore compared NMM fluorescence of Lin28-binding RNAs (pre-let-7a, let-7a loop, let-7g loop, let-7d loop, and the LREs of DSG2, HSPA5, SUN1 and OCT4) with that of RNAs that lack a G4

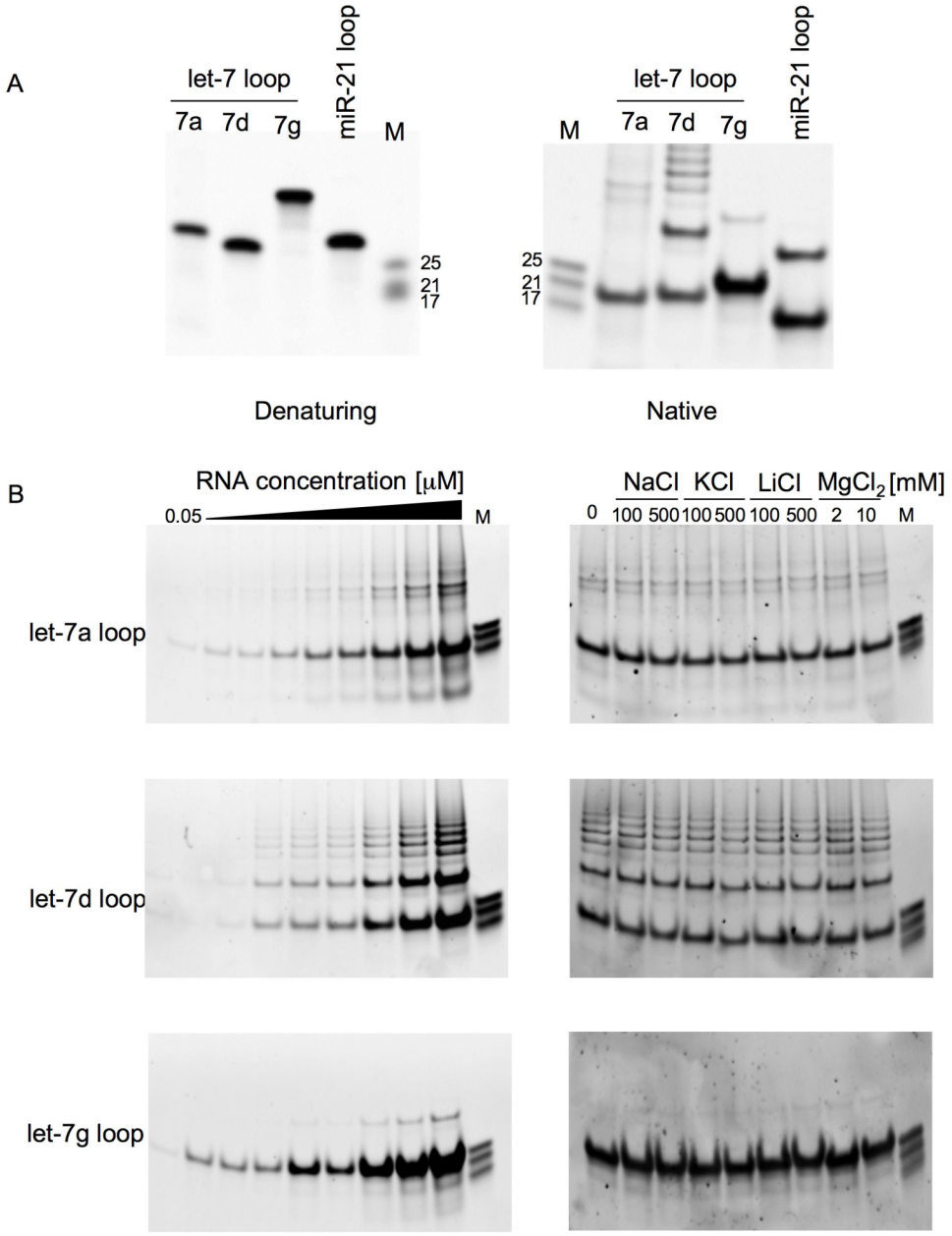


Figure 2.9 Lin28 RNAs migrate as multiple bands. PAGE analysis under denaturing conditions (heat and 8 M Urea) and native conditions (A). In (B) RNAs were analyzed at concentrations ranging from 0.05-10 μM (left) and at 1μM concentration with the specified concentrations of counter ions (right). RNAs of 25, 21 and 17 nt were used as markers (M). RNA was visualized using SYBR-gold staining.

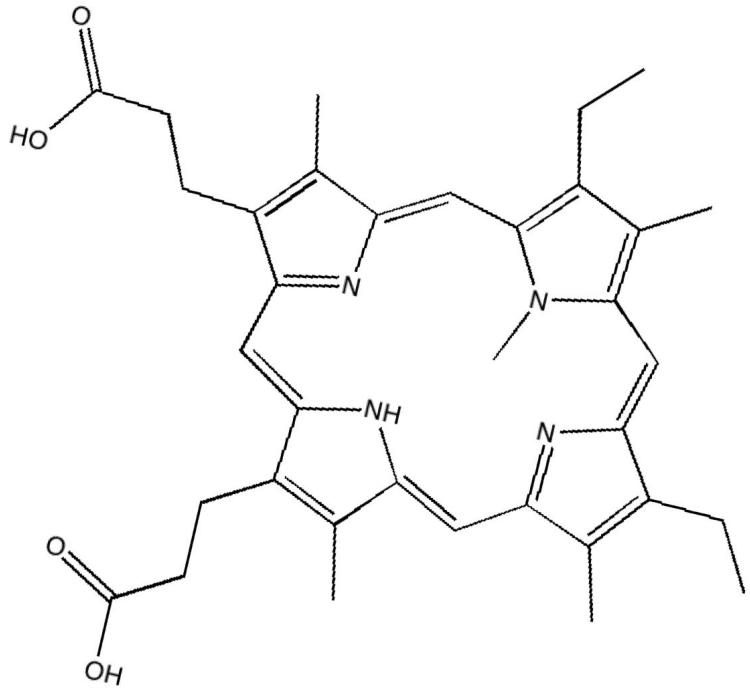


Figure 2.10 N-methyl mesoporphyrin IX structure. N-methyl mesoporphyrin IX (NMM) is a fluorescent dye that selectively binds to the planar structure of G-quartets (14-17).

structure. Strong NMM fluorescence was measured for the known G4 RNA UG4U (UGGGGU) and for all of the Lin28-binding RNAs measured, but little to no fluorescence was detected when NMM was mixed with yeast-tRNA, polyU or the loop of miR-21 (**Figure 2.11**). SUN1, which had weakest affinity for Lin28, based on the Lin28 concentration needed to observe a gel shift (**Figure 2.1**), also had the least NMM fluorescence. Thus Lin28 affinity may correlate with NMM fluorescence. A higher molar concentration of let-7 miRNAs than LRE RNA was needed to saturate NMM fluorescence. Taken together with the 1D NMR results, NMM fluorescence suggests that Lin28-binding miRNAs and mRNAs form G4s.

RNA mutations that disrupt Lin28 binding disrupt NMM fluorescence

To probe further whether Lin28 binding correlates with NMM binding, and presumed G4 formation, we generated RNAs that were truncated or contained mutations that might affect G4 formation. For the let-7g loop we mutated $C_{18}C_{19}$ to $U_{18}U_{19}$ (C-Mt), $G_{27}G_{28}$ to $A_{27}A_{28}$ (G-Mt1) and G_4G_5 and $G_{27}G_{28}$ to A_4A_5 and $A_{27}A_{28}$ (G-Mt2). We also synthesized a HSPA5 LRE truncation by removing 9 nt from the 5'end containing the first two GG-repeats (HSPA5-tr). We tested the ability of each mutant to bind Lin28 by EMSA and to bind to NMM by fluorescence assay (**Figure 2.12**). Compared to the wild-type let-7g-loop, the C-Mt had increased affinity for Lin28 as suggested by the relative intensity of the gel-shifted band with that of free let-7g loop and increased NMM fluorescence. G-Mt1 had reduced Lin28 binding and weaker NMM fluorescence. G-Mt2 did not bind to Lin28 and showed no NMM fluorescence. HSPA5-tr did not cause a Lin28 gel shift or NMM fluorescence, suggesting that G-repeats are critical for

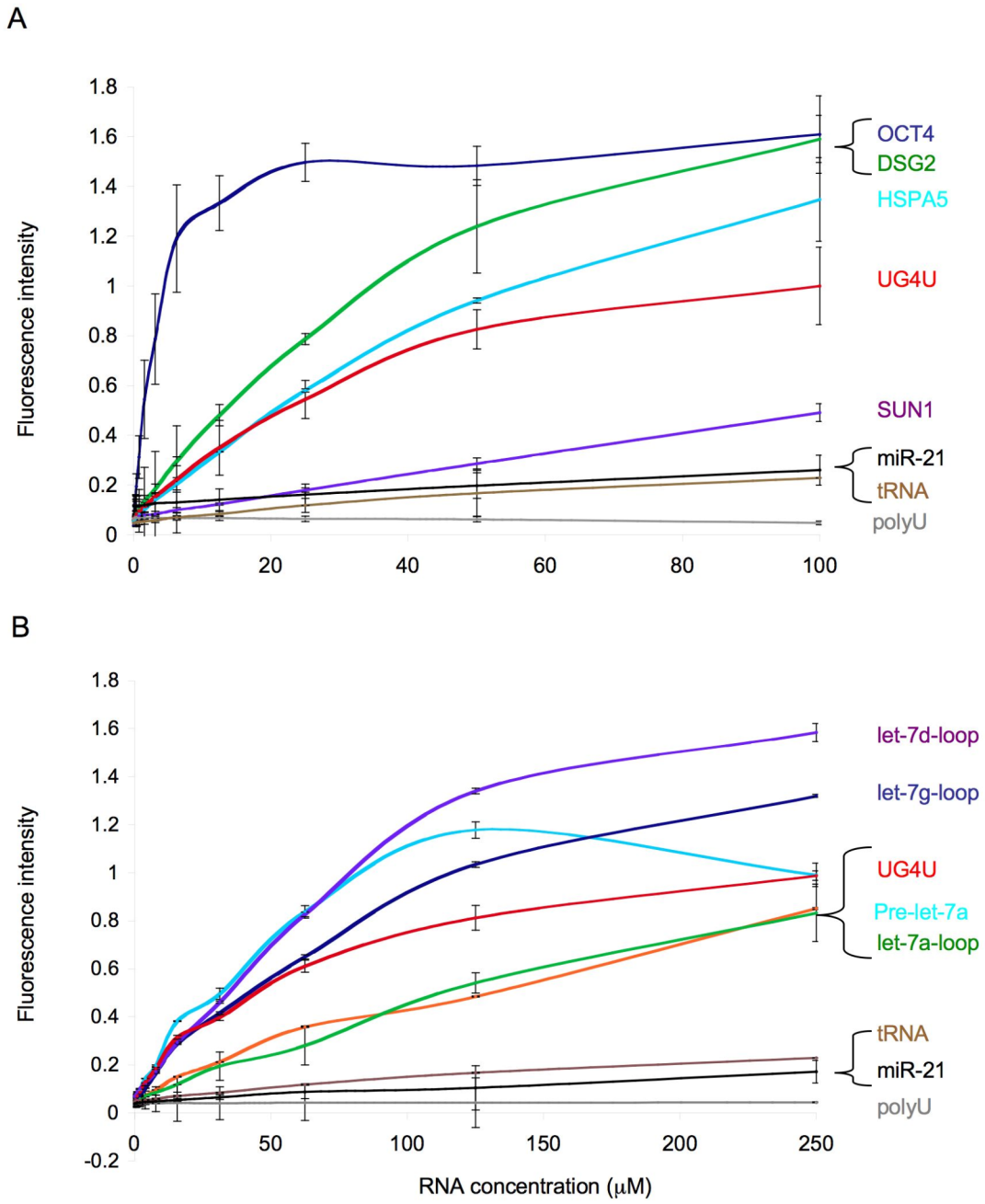


Figure 2.11 Lin28-binding RNAs bind NMM. NMM is a dye that binds G4-containing nucleic acids and then fluoresces at 614 nm (17). NMM was incubated with increasing concentrations of RNA. In (A) are mRNA LREs and (B) depicts let-7 targets. PolyU, tRNA and the miR-21 loop did not bind NMM.

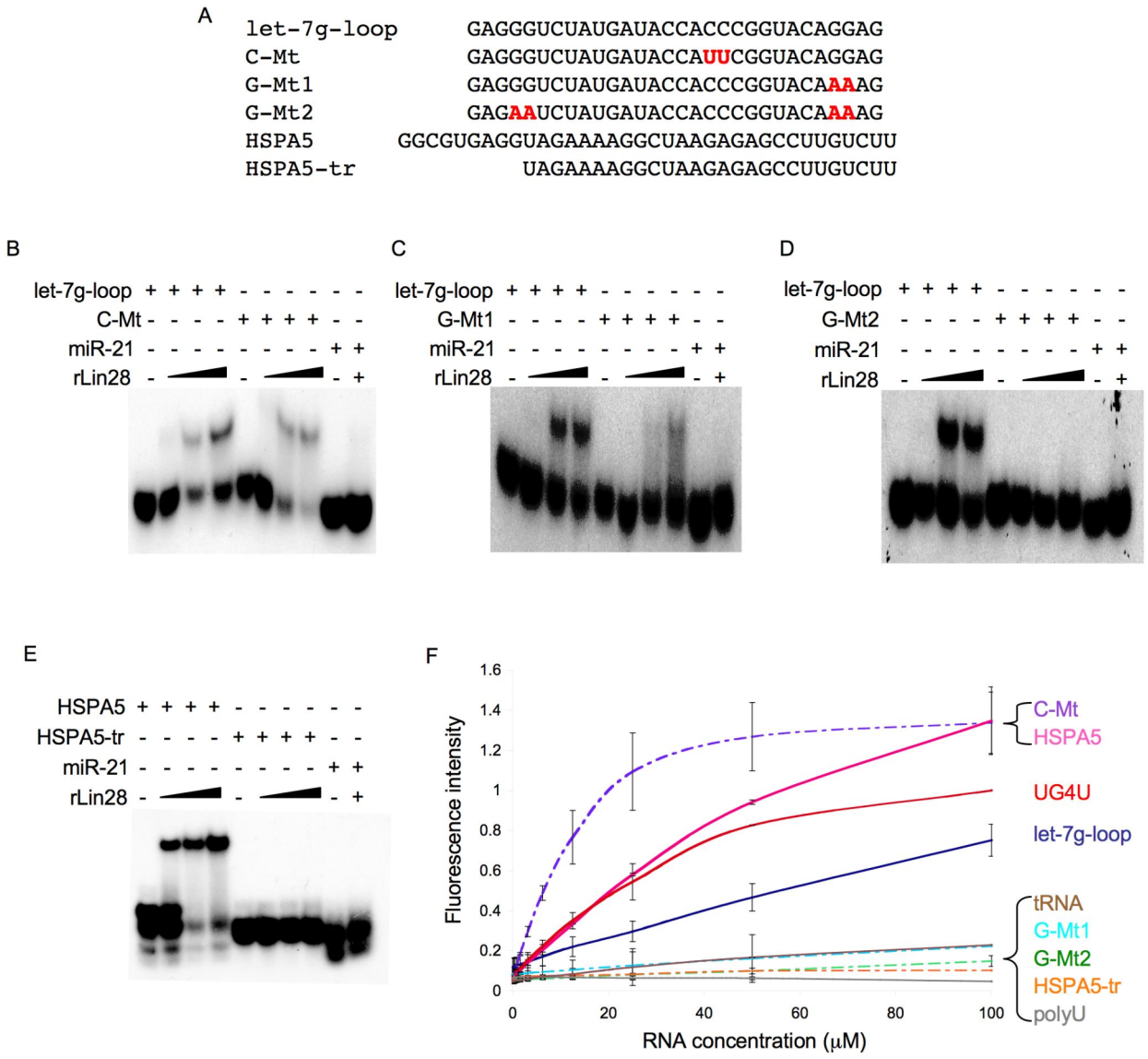


Figure 2.12 Lin28 binding correlates with NMM fluorescence. (A) let-7g loop mutations and HSPA5 truncation evaluated for gel shift and NMM binding. EMSAs comparing the Lin28 interaction with wild-type and mutant let-7g loop (B-D) and full length and truncated HSAP5 LRE (E). NMM fluorescence assay for wild-type and mutant let-7g loop and full length and truncated HSPA5 LRE (F).

Lin28 binding and that Lin28 binding correlates with NMM fluorescence. These findings suggest that Lin28 and NMM recognize a shared structural feature of RNA. Since NMM is only known to bind G-quartets, Lin28 likely recognizes a G4 structure in its targets.

Lin28 mRNA target datasets are enriched for sequences with the potential to form G4s

We next utilized bioinformatics to determine if a G4 structure might be a unifying theme among a larger group of Lin28 RNA targets. QGRS (Quadruplex G-Rich Sequence) Mapper predicts G4s in nucleic acids by identifying potential G-quartet sequences that match the following motif: $G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$, where $_x$ is the number of G-repeats and $_{y1,y2,y3}$ are the length of the gaps (*18*). The program generates a G-score such that sequences with higher scores are more likely to form G4s.

To determine if Lin28 binding RNAs are enriched for sequences with the potential to form G4s we compared G-scores of Lin28 targets including the Peng dataset of Lin28-bound mRNAs identified in human embryonic stem cells that were at least 2.5-fold enriched in a Lin28 immunoprecipitation (IP) as compared to pre-immune control IP (*19*) and the Hafner dataset of Lin28-bound mRNAs that were identified by Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) in FLAG/HA-Lin28 overexpressing HEK 293 cells (*20*) to G-scores of other classes of mRNAs (**Figure 2.13**). FMRP (Fragile-X-mental-retardation protein-1) binds mRNAs that contain G4s (*21-23*).

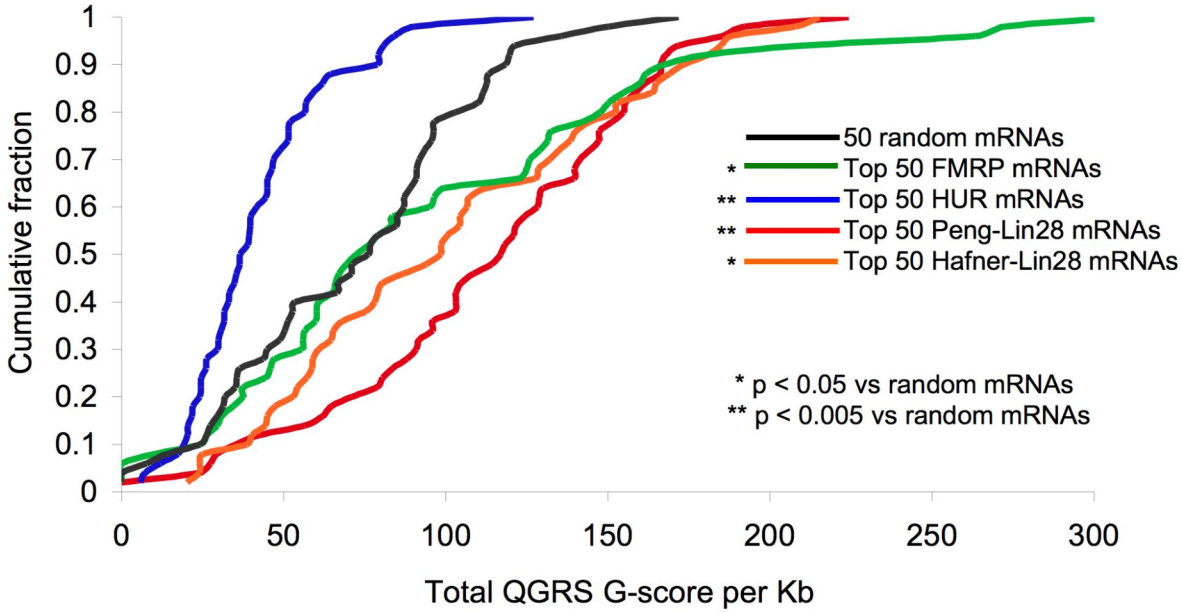


Figure 2.13 Lin28-binding mRNAs are predicted to form G4s. Cumulative distribution plot of G-scores generated by the G4 prediction tool QGRS Mapper (18). Sequences with higher G-scores are more likely to form G4s. FMRP target mRNAs form G4s (20-22) and FMRP mRNA targets (green) have increased G-scores compared to random mRNAs (black) ($p < 0.05$). HUR binds AU-rich mRNAs (23), and HUR mRNA targets (blue) have decreased G-scores compared to random mRNAs ($p < 0.005$). The top 50 enriched Lin28-bound mRNAs (red) in the Peng dataset of Lin28 targets (19) have significantly increased G-scores compared to random mRNAs ($p < 0.005$). The top 50 Lin28 mRNA targets in the Hafner dataset identified by PAR-CLIP (orange) (20) had higher G-scores than random mRNAs ($p < 0.05$). P-values were calculated using a Kolmogorov-Smirnov (K-S) test.

We therefore selected the top 50 enriched FMRP mRNA targets to test the algorithm. The G-scores of HUR target mRNAs were used as a negative control for predicting G4s as HUR is known to bind AU-rich mRNAs (*24*). We also calculated G-scores for 50 random mRNAs. As expected, FMRP-binding mRNAs had significantly higher G-scores than random mRNAs ($p = 0.017$) and the HUR mRNAs had significantly lower G-scores than the random mRNAs ($p < 0.0001$) (**Figure 2.13)**. The top 50 enriched Lin28-bound mRNAs in the Peng dataset (*19*) had significantly higher G-scores than random mRNAs ($p < 0.0001$). The top 50 Lin28 target mRNAs from the Hafner dataset also had higher G-scores than random mRNAs ($p = 0.017$). The top 50 Peng Lin28-binding mRNAs also had significantly higher G-scores than the FMRP target mRNAs ($p = 0.004$). Thus Lin28-binding RNAs are highly enriched for sequences predicted to form G4s. The increased G-score of Lin28 mRNAs from multiple target datasets supports our hypothesis that a common feature among Lin28-binding RNAs might be the ability to form G4s.

Lin28 binds GGA-repetitive G4s

We next evaluated whether Lin28 binds to other G4 containing nucleic acids. UG4U, the thrombin binding aptamer (TBA) and $d(GGA)_8$ all contain G4s that form distinct quadruplex structures (**Figure 2.14**). UG4U is an intermolecular quadruplex formed by 4 separate strands coming together to create 4 stacked G4s in parallel (*25*), TBA (d(GGTTGGTGTGGTTGG)) forms a chair type anti-parallel quaduplex with two stacked G4s (*26*) while $d(GGA)_8$ is unique, in that its structure contains both a G-quartet (G:G:G:G) and a mixed

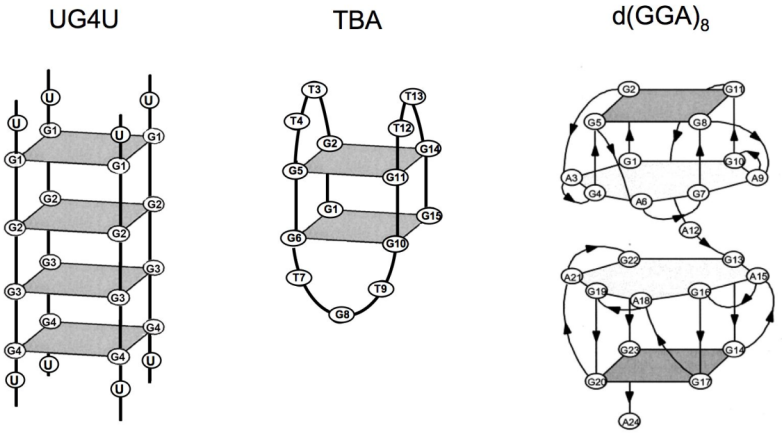


Figure 2.14 G-quadruplexes form diverse structures. Known quadruplexes UG4U with sequence UGGGGU, the thrombin binding aptamer (TBA) with sequence d(GGTTGGTGTGGTTGG) and d(GGA)$_8$ form distinct conformations based on the number of stacked G4s, loop sequences and whether they are formed by inter- or intramolecular G4s.

(G:A:G:A:G:A:G) heptad plane (*27*). GGA-repetitive quadruplexes like $d(GGA)_8$ can also form A(G:G:G:G)-pentad and A(G:G:G:G)A-hexad structures (*28-30*). As expected UG4U, TBA and $d(GGA)_8$ all exhibit strong NMM fluorescence. However, despite NMM binding, only $d(GGA)_8$ bound to Lin28 by EMSA (**Figure 2.15**). Thus, Lin28 does not bind to all nucleic acids that contain G4s. However, we wondered whether a structural feature of $d(GGA)_8$ might facilitate its interaction with Lin28. Sequence comparison of Lin28 binding RNAs with $d(GGA)_8$ demonstrated that multiple GGA-rich tracts are present in Lin28 target sequences, although sometimes the order of As and Gs is reversed. In addition, the 5'terminal 28 nt sequence of the HSPA5 LRE is nearly identical to $d(GGA)_8$ (**Figure 2.16**).

Experimentally validated Lin28-binding RNAs are enriched for G-repeats

For each Lin28 target with an experimentally defined LRE, we calculated the frequency of each base and tabulated the number and length of G-repeats, GGA-repeats and AGG-repeats in each LRE (**Table 2.1**). Most LREs are enriched for guanines (average 30%) and have multiple stretches of 2-6 consecutive guanines and at least one GGA- or AGG-repeat. The frequency of G was not enhanced mostly for the longer LREs (200-3270 nt), which may not correspond to the minimal LRE sequence. G-repeats are under-represented compared to repeats of other bases in exon sequences (*31*) (**Table 2.2**). Although the overall frequency of bases within exons is relatively the same for G, C, A, and T, searching for the pattern ($X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+}$) for X=G, C, A,T and N is any nucleotide showed about half as many G-repeats than repeats of

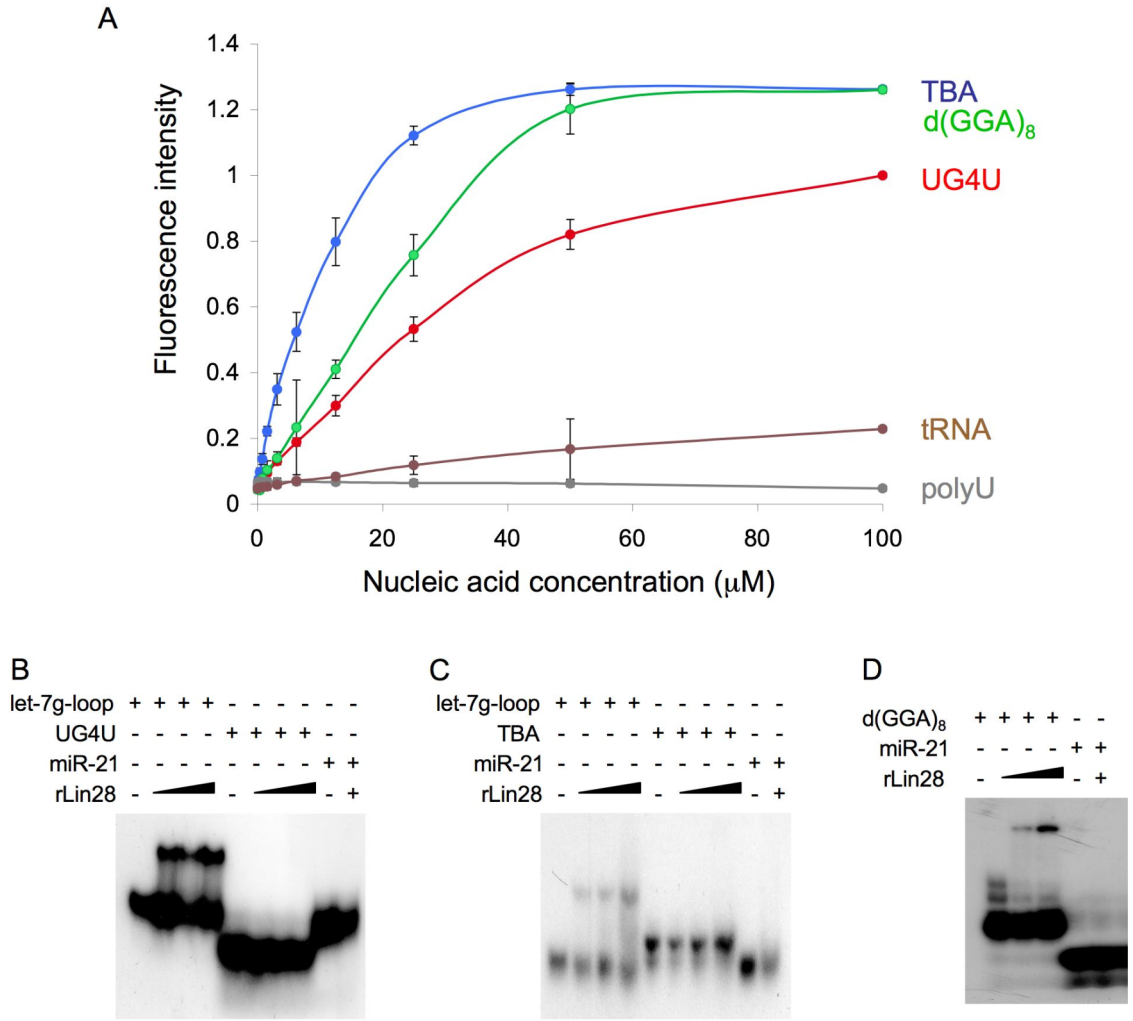


Figure 2.15 Lin28 does not bind all G4 containing nucleic acids. (A) NMM fluorescence assay confirming that UG4U, the thrombin binding aptamer (TBA) and d(GGA)$_8$ form G-quartets. (B-D) EMSAs indicating that Lin28 does not bind to UG4U (B) or the thrombin-binding aptamer (C), but does form a complex with d(GGA)$_8$ (D).

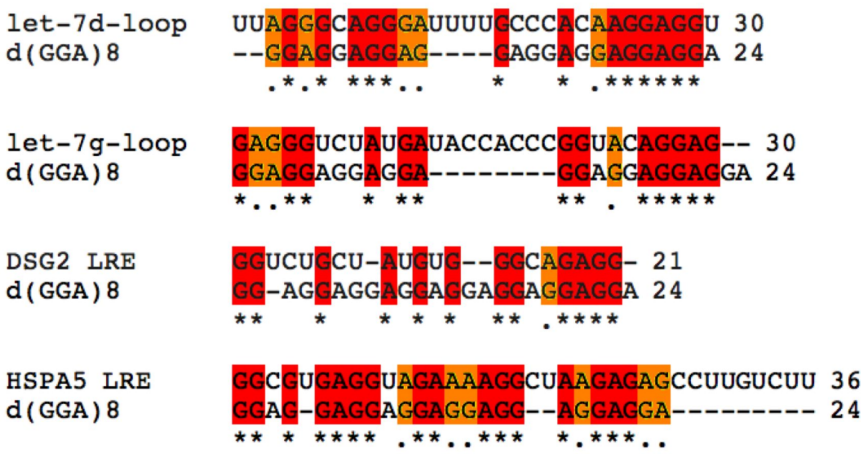


Figure 2.16 Lin28-binding RNAs have sequence similarities to GGA-repetitive G4s. (A) Individual clustalw sequence alignments highlighting the sequence similarity between d(GGA)$_8$ and selected Lin28-binding RNAs. Conserved nucleotides are shown in red and purine substitutions are shown in orange.

Table 2.1 LREs are enriched for guanine (Continued). The frequency of each base (A, U, G, C) was calculated for LREs of Lin28-binding RNAs with a defined LRE. The most enriched nucleotide is shown in red. The number and length of G-repeats, GGA-repeats, and AGG-repeats was also tabulated.

Table 2.1 LREs are enriched for guanine (Continued).

Lin28 binding RNA	%A	%U	%C	%G	GG	GGG	GGGG	G5+	GGA	AGG	Length
IGF-2	13.2	19.9	42.2	24.7	57	15	5	-	15	11	1180
H2A	26.5	10.3	30.1	33.1	9	3	-	-	2	5	136
CyclinB	27.4	39.4	18.6	14.5	8	2	-	-	3	4	317
CDK4	24.9	27.2	28.7	19.2	12	3	1	-	4	6	261
OCT4	24,2	20.0	25.3	30.5	9	4	2	1	3	3	95
EEF1G	27.6	17.6	25.2	29.7	25	3	-	-	14	13	330
RPS13	27.6	26.3	23.0	23.0	19	5	-	-	8	8	456
HMGA1	22.5	7.8	30.2	39.5	14	5	3	-	4	3	129
TDP-43	29.2	23.5	11.8	36.5	7	1	-	-	2	4	85
Lin28	25.0	29.4	22.3	23.3	178	59	16	9G5, 3G6	59	71	3270
DSG2	14.3	23.8	14.3	47.6	3	1	-	-	-	1	21
HSPA5	27.8	22.2	13.9	36.1	3	-	-	-	-	2	36
SUN1	35.0	15.0	15.0	35.0	2	-	-	-	2	1	20
RPS19	17.9	18.9	31.1	32.1	9	2	2	-	-	-	106
HER-2	20.0	16.5	36.5	27.0	14	3	-	-	5	10	200
let-7a-1 loop	28.6	26.8	19.6	25.0	4	3	1	1	1	1	56
let-7a-2 loop	30.0	28.0	14.0	28.0	3	2	1	1	1	1	50
let-7a-3 loop	20.8	32.1	17.0	30.2	5	3	2	1	1	-	53
let-7b loop	20.7	27.6	19.0	32.8	5	2	1	1	1	1	58
let-7c loop	22.0	32.0	16.0	30.0	4	2	1	1	1	-	50
let-7d loop	23.3	23.3	16.7	36.7	4	2	-	-	2	4	30
let-7e loop	26.0	20.0	20.0	34.0	6	1	1	1	3	3	50
let-7f-1 loop	25.0	35.0	11.7	28.3	5	2	2	-	2	1	60
let-7f-2 loop	27.3	32.7	16.4	23.6	4	2	1	-	2	1	55
let-7g loop	26.7	16.7	23.3	33.3	3	1	-	-	1	2	30
let-7i loop	13.3	30.0	18.3	38.3	5	2	1	-	1	0	60
miR-98 loop	29.4	23.5	11.8	35.3	4	2	1	-	1	2	34

Table 2.2 Frequencies of bases and patterns within exons. Frequency of each base in human exonic regions. The observed pattern refers to the frequency of the pattern, $X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+}$ for X = G,C,A,T and normalized to 1 for X=T. Table from Huppert, J.L., et al. (2005).

Base	Frequency	Observed pattern
G	0.25	0.48
C	0.25	0.83
A	0.26	0.93
T	0.24	1.0

other bases (*31*). Evolutionary pressure may have disfavored G-repeats to prevent G4 structures in transcribed RNAs (*31*). Thus the high frequency of G-repeats in Lin28 targets is unusual. Compared to the exon frequency rates, the striking increase of G-repeats within Lin28-binding RNAs suggests that G-repeats are not a result of random chance, but may have evolved to have a functional role.

C. elegans pre-let-7 binds N-methyl mesoporphyrin IX and contains GGA-repeats

Many Lin28-binding RNAs are evolutionarily conserved, as best exemplified by the let-7 family of miRNAs, which are present throughout the animal phyla (*32*). To ascertain if the G4 structure is evolutionarily conserved, we synthesized the *C. elegans* pre-let-7 miRNA (cel-pre-let-7) and tested its ability to bind NMM (**Figure 2.17**). cel-pre-7 induced modest NMM fluorescence, while tRNA, polyU and the miR-21 loop did not. In addition, like other Lin28-binding RNAs, cel-pre-let7 contains several GGA repeats (**Figure 2.17**). Although further studies are required, the ability of cel-pre-let-7 to bind NMM suggests that G4 features might be an evolutionarily conserved structural mechanism for Lin28 recognition.

Lin28 binding disrupts the ability of Lin28 RNA targets to interact with N-methyl mesoporphyrin IX

Upon binding, Lin28 induces a conformational change in its RNA targets (*33-36*). We hypothesized the Lin28-induced RNA rearrangement might unwind the RNA

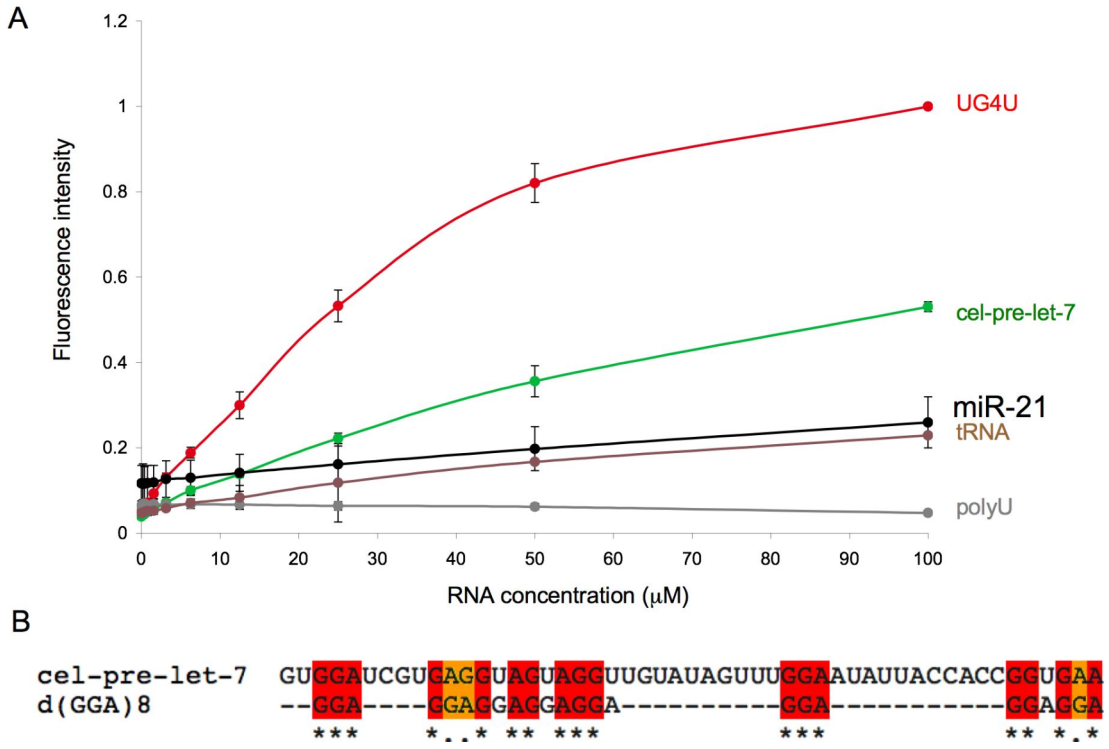


Figure 2.17 *C. elegans* pre-let-7 binds NMM and contains GGA repeats. (A) NMM fluorescence assay indicates that the *C.elegans* pre-let-7 miRNA (cel-pre-let-7) binds NMM, while yeast tRNA, polyU and the miR-21 loop do not. (B) Clustalw sequence alignment of cel-pre-let-7 and d(GGA)$_8$ showing that cel-pre-let-7 contains several GGA repeats. Conserved residues are highlighted in red, purine substitutions are shown in orange.

G4 structure. To test this idea we performed a modified NMM assay in which the let-7g-loop or HSPA5 LRE were incubated with a constant concentration of NMM and increasing amounts of rLin28A or bovine serum albumin (BSA) (**Figure 2.18**) The let-7g-loop and HSPA5 LRE exhibited strong NMM fluorescence on their own, but when Lin28 was added, NMM fluorescence was lost. Addition of BSA had no effect. Thus, in the presence of Lin28, the RNA targets are no longer able to interact with NMM. These results suggest that Lin28 competes with NMM binding or that Lin28 binding unwinds the G4s, abrogating NMM binding.

N-methyl mesoporphyrin IX inhibits Lin28 binding to target RNAs via EMSA

By binding to G4s within gene promoters and within regulatory sites of mRNAs, G4 intercalating agents, like NMM, inhibit transcription and translation *(37-39)*. We performed a competition EMSA to examine whether NMM inhibits Lin28 binding (**Figure 2.19**). Both ^{32}P-labeled let-7g-loop and HSPA5 LRE bound rLin28A as measured by the presence of a gel shift. However, adding increasing amounts of NMM led to a decrease in the gel shift and an increase in the signal of free RNA. Thus NMM inhibits Lin28's ability to bind RNA. To our knowledge this is the first example of a small-molecule inhibitor of Lin28 binding. The miR-21 loop and HSPA5 LRE truncation (HSPA5-tr) did not bind Lin28 or NMM at the highest concentration.

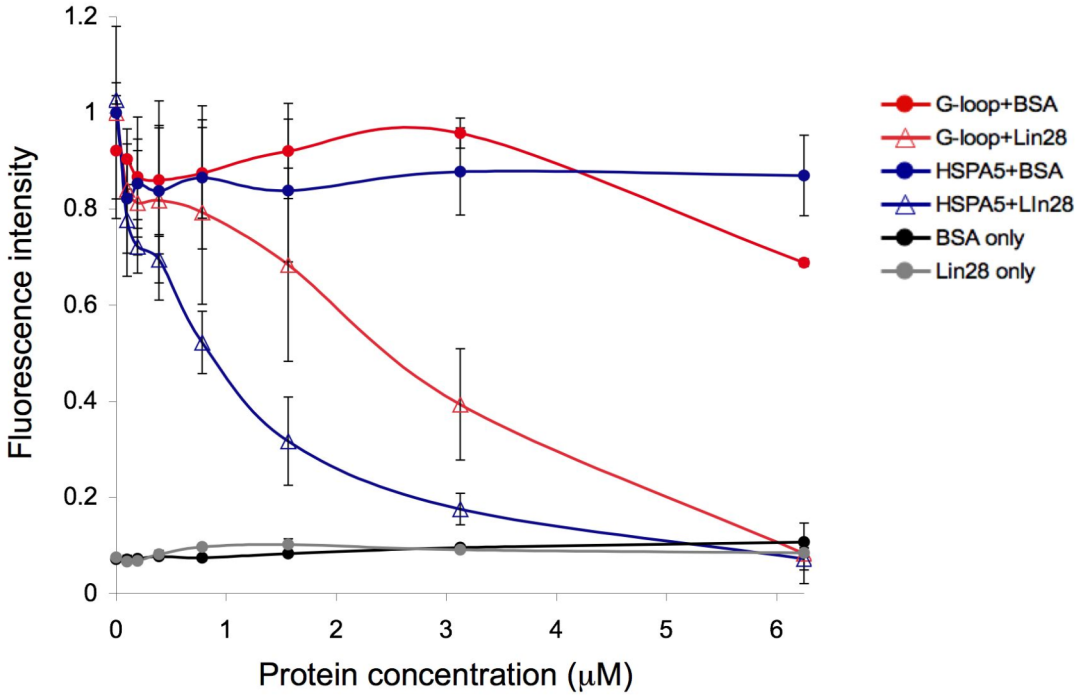


Figure 2.18 Lin28 binding inhibits NMM fluorescence. 50 μM of the let-7g-loop or HSPA5 LRE were incubated with a constant amount of NMM (5 μM). When increasing amounts of rLin28A were added, NMM fluorescence diminished. BSA had no effect. Lin28 or BSA on their own did not fluoresce.

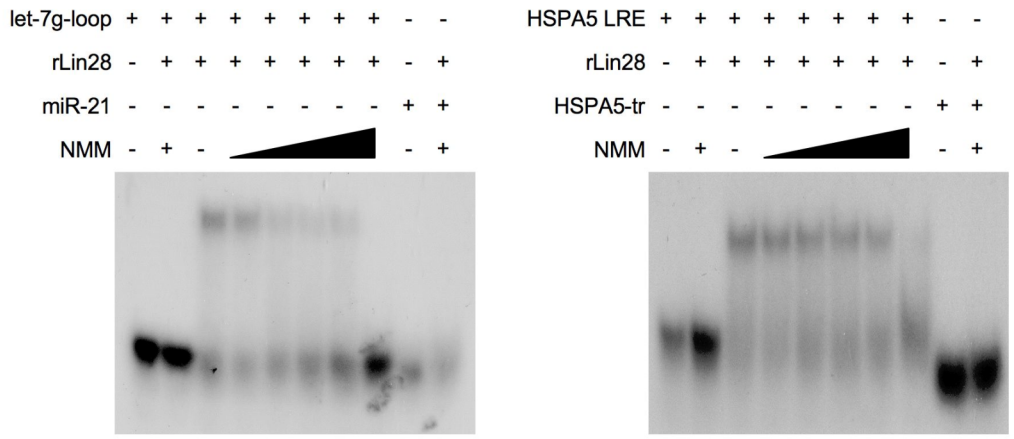


Figure 2.19 NMM inhibits Lin28 binding. Competition EMSAs for Lin28 binding to the let-7g-loop (A) and HSPA5 (B). ^{32}P-labeled RNAs were incubated with a constant amount of rLin28A (10 μM) known to form a Lin28•RNA gel shift. Increasing amounts of NMM (0-100 μM) diminished the amount of rLin28A shifted product and increased signal of the free ^{32}P-RNA. NMM and rLin28A had no effect on miR-21 or HSPA5-tr.

Materials & Methods:

Lin28 expression and purification

The cDNA for mouse Lin28A were obtained from Addgene (Cambridge, MA). The region containing the CSD and ZnFs (residues 36-186) was PCR amplified and cloned into the bacterial expression vector pET15b, adding a C-terminal HIS_6 tag. Lin28-HIS_6 was overpressed in *E. coli* Rosetta (DE3) cells (Millipore). Cells were grown in Luria Broth (LB) medium at 25°C to an OD_{600} of 0.9 and induced with 1.0 mM isopropyl D-1-thiogalactopyranoside (IPTG). After 3 hours, cells were harvested by centrifugation. Cells were lysed by sonication in 50 mM Tris-Cl pH 8.0, 1M NaCl, 5 mM β-mercaptoentanol (BME), 0.1 mM $ZnCl_2$ and 10 units of Benzonase nuclease (Sigma). Lin28-HIS_6 was purified by Ni affinity chromatography followed by size-exclusion chromatography on a Superdex 75 prep-grade column in 50 mM Tris-Cl pH 6.4, 200 mM NaCl, 5 mM BME. All purification steps were carried out on ice or at 4°C.

Lin28-binding RNAs

RNAs were synthesized by Integrated DNA Technologies (Coralville, IA) and obtained as a lyophilized powder. RNAs were reconstituted in 10 mM sodium phosphate buffer, pH 6.4, 100 mM KCl, 4 mM $MgCl_2$. Annealing was carried out by heating the samples to 80°C for 10 minutes and allowing them to cool overnight to 20°C.

RNA sequences:

pre-let-7a-1

UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUCACACCCACCACUGG

GAGAUAACUAUACAAUCUACUGUCUUUCCUA

pre-let-7g

AGGCUGAGGUAGUAGUUUGUACAGUUUGAGGGUCUAUGAUACCACCCGG

UACAGGAGAUAACUGUACAGGCCACUGCCUUGCCA

let-7a-1 loop

GGUUGUAUAGUUUUAGGGUCACACCCACCACUGGGAGAUAACUAUACAAU

CUA

let-7d loop

UUAGGGCAGGGAUUUUGCCCACAAGGAGGU

let-7g loop

GAGGGUCUAUGAUACCACCCGGUACAGGAG

cel-pre-let-7

UACACUGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUAC

CACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAGACAGAACUCUUCGA

DSG2 LRE

GGUCUGCUAUGUGGGCAGAGG

HSPA5 LRE

GGCGUGAGGUAGAAAAGGCUAAGAGAGCCUUGUCUU

SUN1 LRE

UAUCCAGAAGGAGCUGGAAG

OCT4 LRE

GCAGAAGAGGAUCACCCUGGGAUAUACACAGGCCGAUGUGGGGCUCACC

CUGGGGGUUCUAUUUGGGAAGGUAUUCAGCCAAACGACCAUCUGCC

miR-21 loop

GAUGUUGACUGUUGAAUCUCAUGGCAACAC

Electrophoretic mobility shift analysis

EMSAs were conducted using $2x10^5$ cpm ^{32}P-5'-end-labeled RNA probes. Binding buffer contained 50 mM Tris (pH 7.6), 100 mM KCl, 5% glycerol, 40 units of RNAseOUT (Invitrogen) and 10 mM BME. Binding reactions were conducted in 20 μl with increasing amounts of recombinant Lin28-His_6 (0, 0.15, 0.3, 0.6, 1.25, 2.5, 5.0, 10 μM). Bound complexes were resolved on native 5% polyacrylamide gels and visualized by autoradiography.

NMR spectroscopy

NMR spectra were acquired on Bruker 500 MHz or 600 MHz spectrometers equipped with cryoprobes. 1D ^{1}H NM spectra were acquired after annealing 200-500 μM RNA in 10 mM sodium phosphate buffer, pH 6.4, 100 mM KCl, 4 mM $MgCl_2$ in 90%H_2O/10%D_2O. Samples were equilibrated at 5°C. Data was accumulated with 1024 transients over a frequency width of 10.33 kHz with 32K data points. Spectra were processed using MestReNova. 2D NOESY spectra in H_2O were acquired at mixing times from 80-200 ms at temperatures from 10°C °C -40°C. 2048 complex points were acquired in the indirect dimension and 1024 free induction decay (FIDs) were accumulated in the direction dimension. A relaxation delay of 1.5 s was used for all experiments.

N-methylmesoprphyrin IX (NMM) fluorescence assay

NMM was obtained from Frontier Scientific (Logan, UT). Fluorescence assays were performed in 30 μL of 10 mM sodium phosphate buffer, pH 6.4, 100 mM KCl, 4 mM $MgCl_2$ and 5 μM NMM. The RNA concentration ranged from 0-250 μM. All fluorescence experiments were preformed using a FlexStation III plate reader with excitation and emission wavelengths of 399 nm and 614 nm, respectively. Fluorescence measurements were repeated three times for each sample and the intensities averaged and corrected by running a buffer control without RNA before each series of experiments. Fluorescence intensities were normalized to the maximum intensity of UG4U. Results shown are the average of 3-5 independent replicates. Error bars represent the standard deviation between experiments.

G-score cumulative distribution plots

Cumulative distribution plots were calculated using the G-scores per Kb generated by Quadruplex G-rich Sequence (QGRS) Mapper (*18*) for the indicated mRNA target lists. A Kolmogorov-Smirnov (K-S) statistical test (*40*) was used to calculate p-values and determine how the G-score of each mRNA list compares against the random mRNA distribution.

Unwinding assay

Unwinding assays were carried out with constant amounts of let-7g loop or HSPA5 LRE (50 μM) in 10 mM sodium phosphate buffer, pH 6.4, 100 mM KCl, 4 mM $MgCl_2$ and 5 μM NMM to which rLin28A or BSA protein was added. The protein concentration ranged from 0-6.25 μM. After the addition of protein,

samples were allowed to equilibrate at least 4hrs at 4°C. Fluorescence intensities were recorded using a FlexStation III plate reader, as in the NMM fluorescence assay. Protein only samples were recorded in 10 mM sodium phosphate buffer, pH 6.4, 100 mM KCl, 4 mM $MgCl_2$ and 5 μM NMM without RNA. Fluorescence intensities were normalized to the maximum intensity without protein. Results shown are the mean value and standard deviation of 3 replicates.

Competition EMSA

Competition EMSAs were conducted using $2x10^5$ cpm ^{32}P-5'-end-labeled let-7g loop or HSPA5 LRE probes and a constant rLin28A (10 μM). Competition reactions were conducted in the same EMSA binding buffer with increasing amounts of NMM (0, 1.5, 3, 6, 12.5, 25, 50, 100 μM). Bound complexes were resolved on native 5% polyacrylamide gels and visualized using autoradiography.

References:

1. Webba da Silva, M. (2007) NMR methods for studying quadruplex nucleic acids, *Methods 43*, 264-277.

2. Adrian, M., Heddi, B., and Phan, A. T. (2012) NMR spectroscopy of G-quadruplexes, *Methods 57*, 11-24.

3. Ferentz, A. E., and Wagner, G. (2000) NMR spectroscopy: a multifaceted approach to macromolecular structure, *Q Rev Biophys 33*, 29-65.

4. Hennig, M., Williamson, J. R., Brodsky, A. S., and Battiste, J. L. (2001) Recent advances in RNA structure determination by NMR, *Curr Protoc Nucleic Acid Chem 7*, 7.7.1-7.7.30.

5. Roberts, G. C. K. (1993) *NMR of Macromolecules: A Practical Approach*, Oxford University Press, New York.

6. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res 31*, 3406-3415.

7. Lerman, Y. V., Kennedy, S. D., Shankar, N., Parisien, M., Major, F., and Turner, D. H. (2011) NMR structure of a 4 x 4 nucleotide RNA internal loop from an R2 retrotransposon: identification of a three purine-purine sheared pair motif and comparison to MC-SYM predictions, *RNA 17*, 1664-1677.

8. Gellert, M., Lipsett, M. N., and Davies, D. R. (1962) Helix formation by guanylic acid, *Proc Natl Acad Sci USA 48*, 2013-2018.

9. Sen, D., and Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis, *Nature 334*, 364-366.

10. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure, *Nucleic Acids Res 34*, 5402-5415.

11. Patel, D. J., Phan, A. T., and Kuryavyi, V. (2007) Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics, *Nucleic Acids Res 35*, 7429-7455.

12. Lane, A. N., Chaires, J. B., Gray, R. D., and Trent, J. O. (2008) Stability and kinetics of G-quadruplex structures, *Nucleic Acids Res 36*, 5482-5515.

13. Cavanagh, J., Fairbrother, W. J., Palmer, A. G., and Skelton, N. J. (1996) *Protein NMR Spectroscopy Principles and Practice*, Academic Press, San Diego.

14. Nicoludis, J. M., Barrett, S. P., Mergny, J.-L., and Yatsunyk, L. A. (2012) Interaction of human telomeric DNA with N-methyl mesoporphyrin IX, *Nucleic Acids Res 40*, 5432-5447.

15. Nicoludis, J. M., Miller, S. T., Jeffrey, P. D., Barrett, S. P., Rablen, P. R., Lawton, T. J., and Yatsunyk, L. A. (2012) Optimized end-stacking provides specificity of N-methyl mesoporphyrin IX for human telomeric G-quadruplex DNA, *J Am Chem Soc 134*, 20446-20456.

16. Ren, J., and Chaires, J. B. (1999) Sequence and structural selectivity of nucleic acid binding ligands, *Biochemistry 38*, 16067-16075.

17. Arthanari, H., Basu, S., Kawano, T. L., and Bolton, P. H. (1998) Fluorescent dyes specific for quadruplex DNA, *Nucleic Acids Res 26*, 3724-3728.

18. Kikin, O., D'Antonio, L., and Bagga, P. S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences, *Nucleic acids research 34*, W676-682.

19. Peng, S., Chen, L.-L., Lei, X.-X., Yang, L., Lin, H., Carmichael, G. G., and Huang, Y. (2011) Genome-Wide Studies Reveal That Lin28 Enhances the Translation of Genes Important for Growth and Survival of Human Embryonic Stem Cells, *Stem Cells 29*, 496-504.

20. Hafner, M., Max, K. E. A., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013) Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition, *RNA 19*, 1-14.

21. Bole, M., Menon, L., and Mihailescu, M.-R. (2008) Fragile X mental retardation protein recognition of G quadruplex structure per se is sufficient for high affinity binding to RNA, *Mol Biosyst 4*, 1212-1219.

22. Brown, V., Jin, P., Ceman, S., Darnell, J. C., O'Donnell, W. T., Tenenbaum, S. A., Jin, X., Feng, Y., Wilkinson, K. D., Keene, J. D., Darnell, R. B., and Warren, S. T. (2001) Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome, *Cell 107*, 477-487.

23. Darnell, J. C., Jensen, K. B., Jin, P., Brown, V., Warren, S. T., and Darnell, R. B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function, *Cell 107*, 489-499.

24. Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U., and Keene, J. D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability, *Mol Cell 43*, 327-339.

25. Deng, J., Xiong, Y., and Sundaralingam, M. (2001) X-ray analysis of an RNA tetraplex (UGGGGU)(4) with divalent Sr(2+) ions at subatomic resolution (0.61 A), *Proc Natl Acad Sci USA 98*, 13665-13670.

26. Macaya, R. F., Schultze, P., Smith, F. W., Roe, J. A., and Feigon, J. (1993) Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution, *Proc Natl Acad Sci USA 90*, 3745-3749.

27. Matsugami, A., Okuizumi, T., Uesugi, S., and Katahira, M. (2003) Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA, *J Biol Chem 278*, 28147-28153.

28. Kettani, A., Gorin, A., Majumdar, A., Hermann, T., Skripkin, E., Zhao, H., Jones, R., and Patel, D. J. (2000) A dimeric DNA interface stabilized by stacked A.(G.G.G.G).A hexads and coordinated monovalent cations, *J Mol Biol 297*, 627-644.

29. Liu, H., Matsugami, A., Katahira, M., and Uesugi, S. (2002) A dimeric RNA quadruplex architecture comprised of two G:G(:A):G:G(:A) hexads, G:G:G:G tetrads and UUUU loops, *J Mol Biol 322*, 955-970.

30. Zhang, N., Gorin, A., Majumdar, A., Kettani, A., Chernichenko, N., Skripkin, E., and Patel, D. J. (2001) V-shaped scaffold: a new architectural motif identified in an A x (G x G x G x G) pentad-containing dimeric DNA quadruplex involving stacked G(anti) x G(anti) x G(anti) x G(syn) tetrads, *J Mol Biol 311*, 1063-1079.

31. Huppert, J. L., and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome, *Nucleic Acids Res 33*, 2908-2916.

32. Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA, *Nature 408*, 86-89.

33. Loughlin, F. E., Gebert, L. F. R., Towbin, H., Brunschweiger, A., Hall, J., and Allain, F. H.-T. (2011) Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28, *Nature Struct Mol Bio 19*, 84-89.

34. Mayr, F., Schütz, A., Döge, N., and Heinemann, U. (2012) The Lin28 cold-shock domain remodels pre-let-7 microRNA, *Nucleic Acids Res*, 1-15.

35. Desjardins, A., Yang, A., Bouvette, J., Omichinski, J. G., and Legault, P. (2012) Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28, *Nucleic Acids Res 40*, 1767-1777.

36. Lightfoot, H. L., Bugaut, A., Armisen, J., Lehrbach, N. J., Miska, E. A., and Balasubramanian, S. (2011) A LIN28-dependent structural change of pre-let-7g directly inhibits Dicer processing, *Biochemistry 50*, 7514-7521.

37. Grand, C. L., Han, H., Muñoz, R. M., Weitman, S., Von Hoff, D. D., Hurley, L. H., and Bearss, D. J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo, *Mol Cancer Ther 1*, 565-573.

38. Phan, A. T., Kuryavyi, V., Gaw, H. Y., and Patel, D. J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter, *Nat Chem Biol 1*, 167-173.

39. Bugaut, A., Rodriguez, R., Kumari, S., Hsu, S.-T. D., and Balasubramanian, S. (2010) Small molecule-mediated inhibition of translation by targeting a native RNA G-quadruplex, *Org Biomol Chem 8*, 2771-2776.

40. Lopes, R. H. C., Reid, I., and Hobson, P. R. (2007) The two-dimensional Kolmogorov-Smrnov test, in *XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research*, Amsterdam, the Netherlands.

Chapter 3: Conclusions and Future Directions

Lin28 is an evolutionarily conserved RNA-binding protein that regulates cell differentiation (*1, 2*). It has been unclear how Lin28 identifies its target RNAs and distinguishes them from other RNAs. Although multiple RNA sequence motifs have been suggested to influence Lin28 binding to RNA, the motifs are small and abundant in the transcriptome. We hypothesized that Lin28 recognizes a common structural feature in its RNA targets. To address this we compared the RNA structural features of Lin28 binding-RNAs that contain many of the putative Lin28 binding sequence elements, including the pre-let-7 and the pre-let-7 terminal loops with the 3'GGAG motif, the LREs of DSG2, HSPA5 and SUN1, which contain the consensus sequences AAGNNG, AAGNG(N) and (N)UGUG(N), respectively, and the LRE of OCT4 which contains the "A bulge".

By 1D NMR spectroscopy we found that Lin28-binding RNAs contain imino resonances from 10-12.4 ppm, indicative of non-canonical hydrogen bonds (*3-5*). The imino resonance pattern and thermostability of these resonances is characteristic of the G:G:G:G hydrogen bonding networks found in G-quartets (G4s) (*3, 5, 6*). Despite tremendous efforts, we were unable to obtain high-resolution 2D NMR spectra that could provide further structural support. G4-containing RNAs are known to be prone to aggregate (*3, 7-9*), which may have foiled our efforts. In fact on native gels, the Lin28 binding loops and LREs contained multiple bands that could not be eliminated by buffer changes. Future structural studies of Lin28 complexed with a let-7 loop, LRE of an mRNA or GGA-containing G4 such as d(GGA)8 would be worthwhile.

G-quartets (G4s) are structures in which four guanines are interconnected by Hoogsteen base pairing. Parallel-stacked G4s form unique structures known as G-quadruplexes (*3, 10, 11*). Quadruplexes are found within DNAs corresponding to eukaryotic telomeres (*12*), the promoter regions of some oncogenes, including BCL-2 (*13*), c-KIT (*14*), c-MYB (*15*), c-MYC (*16*), HIF1-α (*17*), hTERT (*18*), KRAS (*19*), Rb (*20*), and VEGF (*21*), immunoglobulin gene switch regions (*22*), and ribosomal genes (*23*). Although most studies have focused on DNA-G4s, RNA-G4s are attracting more attention as G4 structures have been identified in HIV RNA (*24*), at pre-mRNA cleavage sites (*25*) and within the UTRs of mRNAs where they influence translation (*26, 27*). An RNA G4 was identified within the Lin28 target IGF2, but outside its LRE (*28*). A structure of a Lin28-binding RNA is required to conclusively determine whether G4s are important determinants for Lin28 binding.

Lin28-binding RNAs interact with the highly specific G4 fluorescent dye, N-methyl mesoporphryin IX (NMM) (*29, 30*). Lin28 RNA targets, but not control RNAs such as the loop of miR-21, induced NMM fluorescence. Mutations and truncations of Lin28 RNA targets that hindered Lin28 binding also led to decreased NMM fluorescence. These results suggest that Lin28 and NMM recognize a similar structure. Because NMM selectively binds only to G4s, Lin28 RNA targets likely contain G4 features that are critical for Lin28 recognition.

Lin28-bound mRNAs are enriched for sequences with increased potential to form G4s. The Peng dataset of Lin28 targets contains 1,200 mRNAs that were enriched at least 2.5 fold in a Lin28 immunoprecipitation (IP), compared to a pre-immune control IP

in human embryonic stem cells (hESCs) (*31*). No significantly enriched sequence was identified within this list of Lin28 targets. However, we found that the top 50 most enriched Lin28 mRNAs from this dataset are highly enriched for sequences likely to form G4s. In addition, the top 50 Lin28 mRNA targets from the Hafner dataset of Lin28-bound mRNAs identified by PAR-CLIP in FLAG/HA-Lin28-overexpressing HEK 293 cells also were predicted to have an increased likelihood to form G4s. The Hafner dataset contains 1,800 Lin28A mRNA targets. Although the authors proposed a AYYHY consensus sequence where Y is a pyrimidine and H is any nucleotide except guanine, they did observe that ~15% of the Lin28 targets were enriched for guanines (*32*). Based on binding assays they proposed Lin28 binds U-rich sequences. However, in every example they tested, the addition of GA-repeats to oligonucleotide sequences enhanced Lin28 binding. The increased potential to form G4s within multiple Lin28 target datasets suggests that the G4 structure might be a unifying feature among all Lin28-binding RNAs. We did not analyze the potential to form G4s of Lin28 mRNA targets identified by the Wilbert and Cho studies (6,000 and 13,000 Lin28 bound mRNAs in human and mouse embryonic stem cells, respectively) (*33, 34*), because neither study ranked the Lin28-binding mRNAs.

G-quadruplexes form diverse structures depending on the number of stacked G4s, strand orientation, loop structures and whether the G4s are formed via inter- or intramolecular interactions (*10, 11*). GGA-repetitive quadruplexes have been identified that form G-quartets and G:A pentads, hexads and heptads (*35-38*). Lin28 binds to $d(GGA)_8$, which forms a mixed G-quartet-G:A heptad structure but not to all G4 nucleic

acids. GGA repeats are present in all of the Lin28 binding sequences and the 5'terminal 28 nts of the LRE of HSPA is nearly identical to $d(GGA)_8$. Moreover the structure of Lin28 bound to pre-let-7g loop sequences identified GGAG as an important motif for binding Lin28 ZnFs domains. These results may suggest that Lin28 recognizes specialized G4s containing GGA-repeats. The NMR data for the let-7a-loop and let-7g-loop contained resonances characteristic of G-quartets and G:A or A:A mismatches. A GGA-repetitive G4 structure would explain these resonances. This type of G4 structure could unify some of the seemingly inconsistence Lin28-binding motifs described in the literature. The 3'GGAG motif of pre-let-7s identified in the crystal structure of Lin28 bound to pre-let-7 loop sequences could form essential components of the GGA-G4. Moreover the As and Gs enriched in the various Lin28 consensus sequences (GAGGA, AAGNNG, AAGNG(N), and (N)UGUG(N), where N is any nucleotide) and the proposed "A bulge" flanked by two G:Cs could be critical residues for the GGA-G4 type architecture. Future structural studies and binding assays with Lin28 and additional GGA-G4s will determine if Lin28 selectively binds GGA-G4s.

Lin28 inhibits let-7 processing across the animal phylum (*39*). Unlike mammalian let-7s, the terminal loop of *C. elegans* pre-let-7 lacks a 3'GGAG motif (*39*). Nucleotides that partially extend into the *C. elegans* let-7 stem bind to the ZnF domains of human Lin28 (*40*). Here we showed that the *C. elegans* pre-let-7, which contains several GGA repeats, binds to NMM, suggesting it may form a G4 structure. This suggests that a G4 structure might be an evolutionarily conserved mechanism for Lin28 recognition.

Recent atomic structures of Lin28 in complex with fragments of pre-let-7 loop sequences did not detect G4s (*40, 41*). However, the structure of the free RNA may not be the same as that bound to Lin28. Footprinting assays of pre-let-7g revealed that Lin28 binding induces a conformational change within the RNA (*42*). Isolated studies with the ZnF domains suggest they partially unfold pre-let-7 targets (*43*). Additionally, a FRET study revealed that the cold-shock domain (CSD) of Lin28 remodels the terminal loop of pre-let-7 by unwinding the upper stem region (*44*). We propose that Lin28 "unwinding" RNA disrupts the G4 structure, since in the presence of Lin28 the Lin28-binding RNAs no longer induce NMM fluorescence.

Unwinding G4s could be essential for Lin28 function. During let-7 biogenesis, Lin28 recruits TUT4 to pre-let-7 transcripts and then TUT4 uridylates pre-let-7s to promote their degradation (*45, 46*). TUT4 does not bind to pre-let-7 in the absence of Lin28 (*46*). Unwinding of the pre-let-7 G4 by Lin28 could be required for the TUT4 interaction and subsequent uridylation (**Figure 3.1**). Since the G4 structure might interfere with ribosome scanning, unwinding G4s in mRNAs could also explain Lin28's ability to stimulate mRNA translation.

G4 intercalating agents have been assessed for anti-cancer therapy (*47-49*). TMPyP4, a porphyrin analog of NMM that binds G4s and double-stranded DNA, blocks c-MYC transcription and decreases tumor growth and increases survival in tumor xenograft studies (*47, 50*). If unwinding G4s is crucial to Lin28 function, the G4 feature

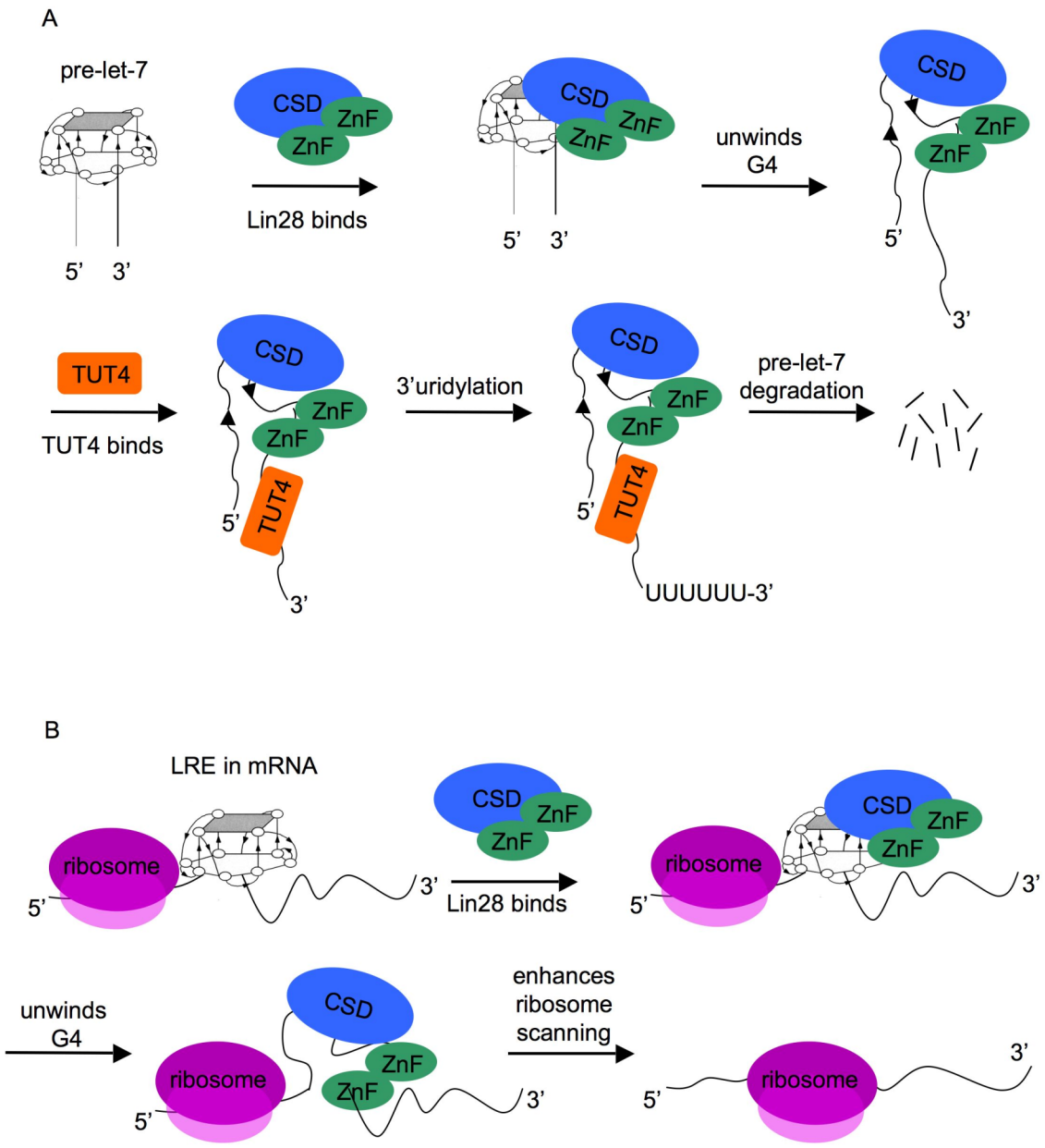


Figure 3.1 Model for Lin28 RNA regulation. We propose that Lin28 binds and unwinds G4s in pre-let-7 miRNAs (A) and in LREs of mRNAs (B). Unwinding pre-let-7s might be required for TUT4 binding, subsequent uridylation and degradation of pre-let-7s. Unwinding G4s in LREs may increase ribosome scanning and stimulate mRNA translation. The Lin28 CSD is in blue and the ZnFs are in green.

of Lin28-binding RNAs could represent a novel site to manipulate Lin28 activity. Here, we have shown that NMM inhibits Lin28 binding to RNAs in a gel shift. This is the first example of a small molecule inhibitor of Lin28. Future studies should investigate whether NMM or similar G4 intercalators can inhibit Lin28 function in cells. Also, since $d(GGA)_8$ is a DNA G4, it is worth investigating whether Lin28 might also recognize and regulate DNA G4-containing sequences.

Summary of This Work:

The work in Part 1 of this thesis identifies a previously unknown common structural feature of Lin28-regulated miRNAs and mRNAs. By NMR spectroscopy, fluorescence assays and bioinformatics we show that the pre-let-7s and the Lin28 responsive elements (LREs) of mRNAs contain G-quartet features (G4s). We propose G4s are unifying features among Lin28-binding RNAs. Our data suggest that Lin28 may specifically recognize GGA-containing planar G4-like structures. Lin28 binding unwinds G4s in its RNA targets. Our findings add new insight to understanding how Lin28 selects and regulates its miRNA and mRNA targets.

References:

1. Moss, E. G., and Tang, L. (2003) Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites, *Dev Biol 258*, 432-442.

2. Viswanathan, S. R., and Daley, G. Q. (2010) Lin28: A microRNA regulator with a macro role, *Cell 140*, 445-449.

3. Adrian, M., Heddi, B., and Phan, A. T. (2012) NMR spectroscopy of G-quadruplexes, *Methods 57*, 11-24.

4. Fürtig, B., Richter, C., Wöhnert, J., and Schwalbe, H. (2003) NMR spectroscopy of RNA, *ChemBioChem 4*, 936-962.

5. Webba da Silva, M. (2007) NMR methods for studying quadruplex nucleic acids, *Methods 43*, 264-277.

6. Lane, A. N., Chaires, J. B., Gray, R. D., and Trent, J. O. (2008) Stability and kinetics of G-quadruplex structures, *Nucleic Acids Res 36*, 5482-5515.

7. Smargiasso, N., Rosu, F., Hsia, W., Colson, P., Baker, E. S., Bowers, M. T., De Pauw, E., and Gabelica, V. (2008) G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation, *J Am Chem Soc 130*, 10208-10216.

8. Marsh, T. C., and Henderson, E. (1994) G-wires: self-assembly of a telomeric oligonucleotide, d(GGGGTTGGGG), into large superstructures, *Biochemistry 33*, 10718-10724.

9. Krishnan-Ghosh, Y., Liu, D., and Balasubramanian, S. (2004) Formation of an interlocked quadruplex dimer by d(GGGT), *J Am Chem Soc 126*, 11009-11016.

10. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure, *Nucleic Acids Res 34*, 5402-5415.

11. Patel, D. J., Phan, A. T., and Kuryavyi, V. (2007) Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics, *Nucleic Acids Res 35*, 7429-7455.

12. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I., and Blackburn, E. H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs, *Cell 51*, 899-908.

13. Dai, J., Dexheimer, T. S., Chen, D., Carver, M., Ambrus, A., Jones, R. A., and Yang, D. (2006) An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution, *J Am Chem Soc 128*, 1096-1098.

14. Fernando, H., Reszka, A. P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A. R., Neidle, S., and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene, *Biochemistry 45*, 7854-7860.

15. Palumbo, S. L., Memmott, R. M., Uribe, D. J., Krotova-Khan, Y., Hurley, L. H., and Ebbinghaus, S. W. (2008) A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity, *Nucleic Acids Research 36*, 1755-1769.

16. Simonsson, T., Pecinka, P., and Kubista, M. (1998) DNA tetraplex formation in the control region of c-myc, *Nucleic Acids Res 26*, 1167-1172.

17. De Armond, R., Wood, S., Sun, D., Hurley, L. H., and Ebbinghaus, S. W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter, *Biochemistry 44*, 16341-16350.

18. Palumbo, S. L., Ebbinghaus, S. W., and Hurley, L. H. (2009) Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands, *J Am Chem Soc 131*, 10878-10891.

19. Cogoi, S., and Xodo, L. E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription, *Nucleic Acids Research 34*, 2536-2549.

20. Xu, Y., and Sugiyama, H. (2006) Formation of the G-quadruplex and i-motif structures in retinoblastoma susceptibility genes (Rb), *Nucleic Acids Research 34*, 949-954.

21. Sun, D., Guo, K., Rusche, J. J., and Hurley, L. H. (2005) Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents, *Nucleic Acids Res 33*, 6070-6080.

22. Maizels, N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells, *Nature Struct Mol Bio 13*, 1055-1059.

23. Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C. B., Proffitt, C., Trent, K., Whitten, J. P., Lim, J. K. C., Von Hoff, D., Anderes, K., and Rice, W. G. (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis, *Cancer Res 69*, 7653-7661.

24. Sundquist, W. I., and Heaphy, S. (1993) Evidence for interstrand quadruplex formation in the dimerization of human immunodeficiency virus 1 genomic RNA, *Proc Natl Acad Sci USA 90*, 3393-3397.

25. Kostadinov, R., Malhotra, N., Viotti, M., Shine, R., D'Antonio, L., and Bagga, P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences, in *Nucleic Acids Res*, pp D119-124.

26. Horsburgh, B. C., Kollmus, H., Hauser, H., and Coen, D. M. (1996) Translational recoding induced by G-rich mRNA sequences that form unusual structures, *Cell 86*, 949-959.

27. Kumari, S., Bugaut, A., Huppert, J. L., and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation, *Nat Chem Biol 3*, 218-221.

28. Christiansen, J., Kofod, M., and Nielsen, F. C. (1994) A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA, *Nucleic Acids Res 22*, 5709-5716.

29. Ren, J., and Chaires, J. B. (1999) Sequence and structural selectivity of nucleic acid binding ligands, *Biochemistry 38*, 16067-16075.

30. Arthanari, H., Basu, S., Kawano, T. L., and Bolton, P. H. (1998) Fluorescent dyes specific for quadruplex DNA, *Nucleic Acids Res 26*, 3724-3728.

31. Peng, S., Chen, L.-L., Lei, X.-X., Yang, L., Lin, H., Carmichael, G. G., and Huang, Y. (2011) Genome-Wide Studies Reveal That Lin28 Enhances the Translation of Genes Important for Growth and Survival of Human Embryonic Stem Cells, *Stem Cells 29*, 496-504.

32. Hafner, M., Max, K. E. A., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013) Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition, *RNA 19*, 1-14.

33. Wilbert, M. L., Huelga, S. C., Kapeli, K., Stark, T. J., Liang, T. Y., Chen, S. X., Yan, B. Y., Nathanson, J. L., Hutt, K. R., Lovci, M. T., Kazan, H., Vu, A. Q., Massirer, K. B., Morris, Q., Hoon, S., and Yeo, G. W. (2012) LIN28 Binds

Messenger RNAs at GGAGA Motifs and Regulates Splicing Factor Abundance, *Mol Cell 48*, 1-12.

34. Cho, J., Chang, H., Kwon, S. C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y. K., and Kim, V. N. (2012) LIN28A Is a Suppressor of ER-Associated Translation in Embryonic Stem Cells, *Cell 151*, 765-777.

35. Zhang, N., Gorin, A., Majumdar, A., Kettani, A., Chernichenko, N., Skripkin, E., and Patel, D. J. (2001) V-shaped scaffold: a new architectural motif identified in an A x (G x G x G x G) pentad-containing dimeric DNA quadruplex involving stacked G(anti) x G(anti) x G(anti) x G(syn) tetrads, *J Mol Biol 311*, 1063-1079.

36. Kettani, A., Gorin, A., Majumdar, A., Hermann, T., Skripkin, E., Zhao, H., Jones, R., and Patel, D. J. (2000) A dimeric DNA interface stabilized by stacked A.(G.G.G.G).A hexads and coordinated monovalent cations, *J Mol Biol 297*, 627-644.

37. Liu, H., Matsugami, A., Katahira, M., and Uesugi, S. (2002) A dimeric RNA quadruplex architecture comprised of two G:G(:A):G:G(:A) hexads, G:G:G:G tetrads and UUUU loops, *J Mol Biol 322*, 955-970.

38. Matsugami, A., Ouhashi, K., Kanagawa, M., Liu, H., Kanagawa, S., Uesugi, S., and Katahira, M. (2001) An intramolecular quadruplex of (GGA)(4) triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad, and its dimeric interaction, *J Mol Biol 313*, 255-269.

39. Lehrbach, N. J., Armisen, J., Lightfoot, H. L., Murfitt, K. J., Bugaut, A., Balasubramanian, S., and Miska, E. A. (2009) LIN-28 and the poly(U) polymerase PUP-2 regulate let-7 microRNA processing in Caenorhabditis elegans, *Nat Struct Mol Biol 16*, 1016-1020.

40. Loughlin, F. E., Gebert, L. F. R., Towbin, H., Brunschweiger, A., Hall, J., and Allain, F. H.-T. (2011) Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28, *Nature Struct Mol Bio 19*, 84-89.

41. Nam, Y., Chen, C., Gregory, Richard I., Chou, James J., and Sliz, P. (2011) Molecular Basis for Interaction of let-7 MicroRNAs with Lin28, *Cell 147*, 1080-1091.

42. Lightfoot, H. L., Bugaut, A., Armisen, J., Lehrbach, N. J., Miska, E. A., and Balasubramanian, S. (2011) A LIN28-dependent structural change of pre-let-7g directly inhibits Dicer processing, *Biochemistry 50*, 7514-7521.

43. Desjardins, A., Yang, A., Bouvette, J., Omichinski, J. G., and Legault, P. (2012) Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28, *Nucleic Acids Res 40*, 1767-1777.

44. Mayr, F., Schütz, A., Döge, N., and Heinemann, U. (2012) The Lin28 cold-shock domain remodels pre-let-7 microRNA, *Nucleic Acids Res*, 1-15.

45. Hagan, J. P., Piskounova, E., and Gregory, R. I. (2009) Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells, *Nat Struct Mol Biol 16*, 1021-1025.

46. Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., and Kim, V. N. (2009) TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation, *Cell 138*, 696-708.

47. Grand, C. L., Han, H., Muñoz, R. M., Weitman, S., Von Hoff, D. D., Hurley, L. H., and Bearss, D. J. (2002) The cationic porphyrin TMPyP4 down-regulates c-MYC and human telomerase reverse transcriptase expression and inhibits tumor growth in vivo, *Mol Cancer Ther 1*, 565-573.

48. Mikami-Terao, Y., Akiyama, M., Yuza, Y., Yanagisawa, T., Yamada, O., and Yamada, H. (2008) Antitumor activity of G-quadruplex-interactive agent TMPyP4 in K562 leukemic cells, *Cancer Lett 261*, 226-234.

49. Murat, P., Singh, Y., and Defrancq, E. (2011) Methods for investigating G-quadruplex DNA/ligand interactions, *Chem Soc Rev 40*, 5293-5307.

50. Phan, A. T., Kuryavyi, V., Gaw, H. Y., and Patel, D. J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter, *Nat Chem Biol 1*, 167-173.

Part 2: Using NMR as a tool for comparing differences in cellular metabolism

Chapter 4: N-acetylneuraminic acid is increased in tumor initiating cells and cytidine monophosphate N-acetylneuraminic acid synthetase is critical for tumor initiation

Attributions

This work was done in close collaboration with Dr. Greg Idos, a postdoctoral fellow in the Lieberman lab. I performed all the NMR experiments (**Figures 4.1 - 4.7**). I performed the mass spectrometry experiments with Colin Hill, a research associate at Agios Pharmaceuticals (**Figure 4.7**). I assisted Dr. Idos with the invasion assays (**Figures 4.10, 4.13, 4.14**). Dr. Idos performed the *in vivo* tumor studies (**Figure 4.16**).

Abstract

Using an unbiased NMR screening method, we sought to identify glucose metabolites that are differentially produced in the highly malignant breast cancer BPLER cell line, compared to the isogenic, but less aggressive, HMLER cell line. Using multi-dimensional NMR experiments, and mass spectrometry we found that N-acetylneuraminic acid (Neu5Ac) is ~7-fold more abundant in BPLER cells. Neu5Ac is added as the terminal sugar to mucinous cell surface glycoproteins. Loss of Neu5Ac by enzymatic removal or siRNA knockdown of the Neu5Ac biosynthetic enzymes N-acetylneuraminic acid synthase (NANS) and cytidine monophosphate N-acetylneuraminic acid synthetase (CMAS) decreased the ability of BPLER cells to invade through a basement membrane. CMAS protein is much more abundant in BPLER than HMLER. Overexpressing CMAS in HMLER cells increases their malignant behavior. BPLER cells stably knocked down for CMAS no longer form tumors in mice. Thus, Neu5Ac is a glucose metabolite that is selectively enriched in BPLER and the expression of Neu5Ac biosynthetic enzyme CMAS is essential for BPLER's ability to initiate tumors.

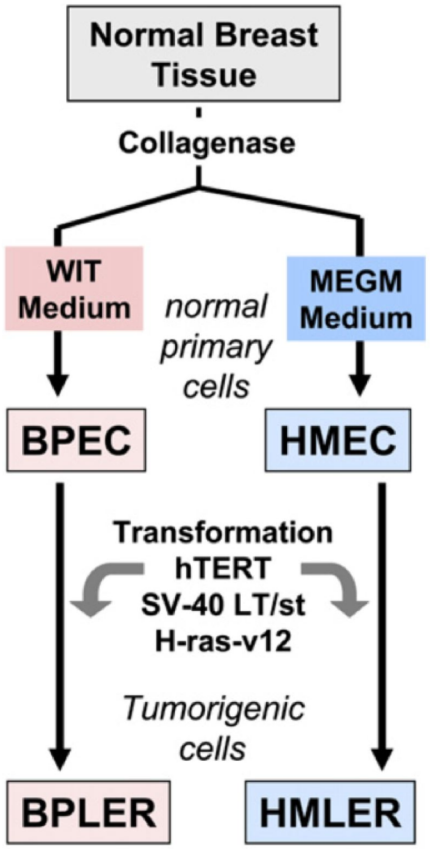
Introduction

Metabolic reprogramming is essential for many cancers. To maintain a high proliferation rate, cancer cells must take up and metabolize nutrients from the environment to produce sufficient ATP to fuel cellular processes and for rapid proliferation (*1, 2*). Complete oxidation of glucose into carbon dioxide via the oxidatitive

phosphorylation pathway generates ~36 molecules of ATP per glucose molecule (*3*). Cancer cells often do not rely on the oxidative phosphorylation pathway, but rather switch to glycolysis, the so called “the Warburg effect” (*4*). Glycolysis generates lactate from glucose and produces only two molecules of ATP per glucose molecule (*3*). Rather than maximizing ATP output, cancer cells divert part of their glucose catabolism to biomass production (*2, 5*). How glucose usage within cancer cells differs from that of normal cells is incompletely defined.

We used an unbiased NMR metabolite-profiling platform to follow the metabolic fate of glucose, comparing a highly malignant human triple negative breast cancer (TNBC) cell line (BPLER) with a less malignant isogenic line (HMLER) generated from the same donor. BPLER and HMLER cells were derived from the same primary breast tissue and transformed with identical genetic factors (hTERT, SV40 LT and $hRAS^{v12}$) (**Figure 4.1)** (*6*). BPLER cells were cultured in chemically defined WIT media and HMLER cells were cultured in MGEM media. BPLER cells have a high tumor initiating capacity. Only 100 BPLER cells formed tumors after subcutaneously injected into the flanks of immunodeficient mice (*6*). However $1x10^6$ HMLER cells are needed to form tumors (*6*). Based on their dramatic differences in malignancy we hypothesized that BPLER cells might utilize glucose differently. We compared glucose metabolites in BPLER cells and HMLER cells as a way to uncover pathways that might be important for tumor initiation or malignancy of TNBC cells.

A



B

number of mice with tumors

Cells injected	BPLER	HMLER
10^6	9/9	5/9
10^5	9/9	0/12
10^4	9/9	0/12
10^3	9/9	0/12
10^2	10/12	-

Figure 4.1 BPLER are more malignant than isogenic HMLER cells. (A) BPLER and HMLER cell lines were derived from the same normal breast tissue, cultured in different medium (WIT or MGEM) and transformed with identical factors (hTERT, SV-40 LT and H-ras-v12). Figure from Ince, T., et al (2007). (B) BPLER cells are more malignant than HMLER cells. Only 100 BPLER cells injected subcutaneously into the flanks of immunocompromised mice formed tumors, whereas $1x10^6$ HMLER cells were needed to form tumors (6).

Results

Highly tumorigenic BPLER cells and less aggressive HMLER cells differentially catabolize glucose.

To compare glucose metabolites generated in BPLER and HMLER cells, we used unbiased NMR metabolite profiling (**Figure 4.2**). BPLER and HMLER cells, cultured in WIT media, were transferred to glucose-free media for ~30 mins before adding uniformly labeled ^{13}C-glucose. Cells were incubated for ~2 hours to allow glucose uptake and catabolism. To assess the ^{13}C background, the spectra of equal numbers of cells that were not incubated with ^{13}C-glucose were also measured. For each cell line, water-soluble metabolites were isolated using chloroform extraction and ultracentrifugation. The aqueous layer was collected and lyophilized. Water-soluble metabolites from ~15 million cells were resuspended in ultra-pure D_2O. For referencing in the NMR analysis, 1mM of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) was added to each sample.

Glucose metabolites were analyzed by ^{13}C-^{1}H heteronuclear single quantum coherence spectroscopy (HSQC). Resonances in HSQC spectra are detected for all ^{13}C-nuclei that are attached to a unique proton (*7*). The HSQC spectra for BPLER and HMLER cells cultured with ^{13}C-glucose had several well-resolved intense peaks (**Figure 4.3**). Replicate samples showed identical resonances. No resonances were observed in the HSQC spectra of unlabeled samples, suggesting that at this low cellular concentration there is little ^{13}C-background. Thus, the ^{13}C-signal in the HSQC spectra from BPLER and HMLER cells incubated with ^{13}C-glucose are metabolite resonances

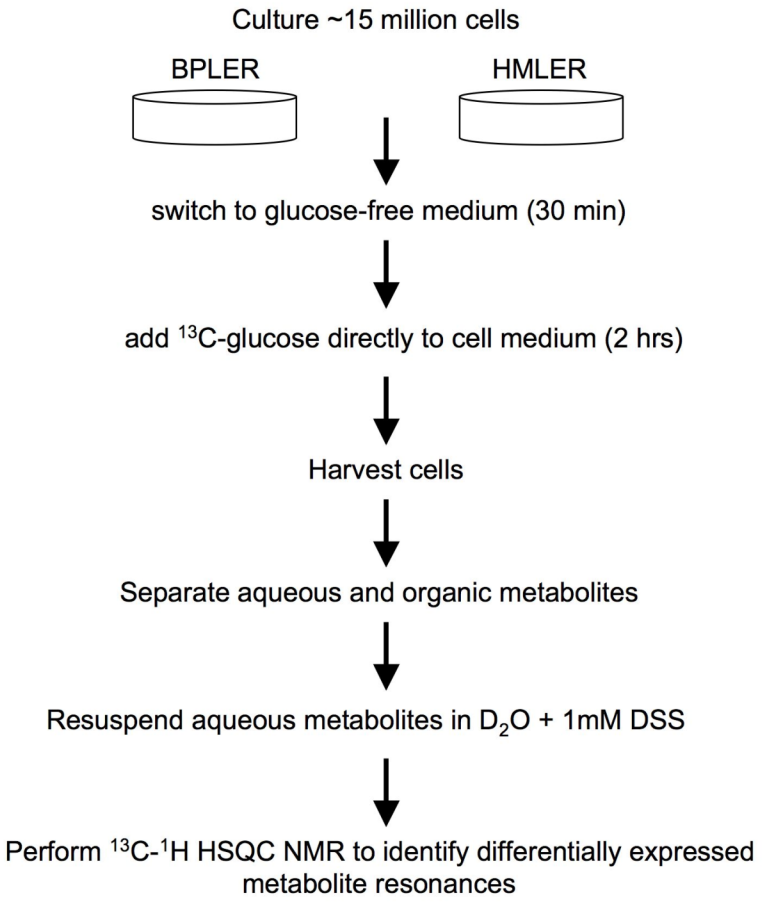


Figure 4.2 Unbiased NMR metabolite profiling workflow. Equal numbers of BPLER and HMLER cells were grown in WIT medium then incubated in glucose-free media. After ~30 min, ^{13}C-glucose was added to to the media and cells were incubated for ~2 hrs. Aqueous and organic metabolites were separated. Water-soluble metabolites were resuspended in D_2O with 1mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS). Differential metabolites were identified by comparing resonances in BPLER and HMLER ^{13}C-^{1}H HSQC spectra.

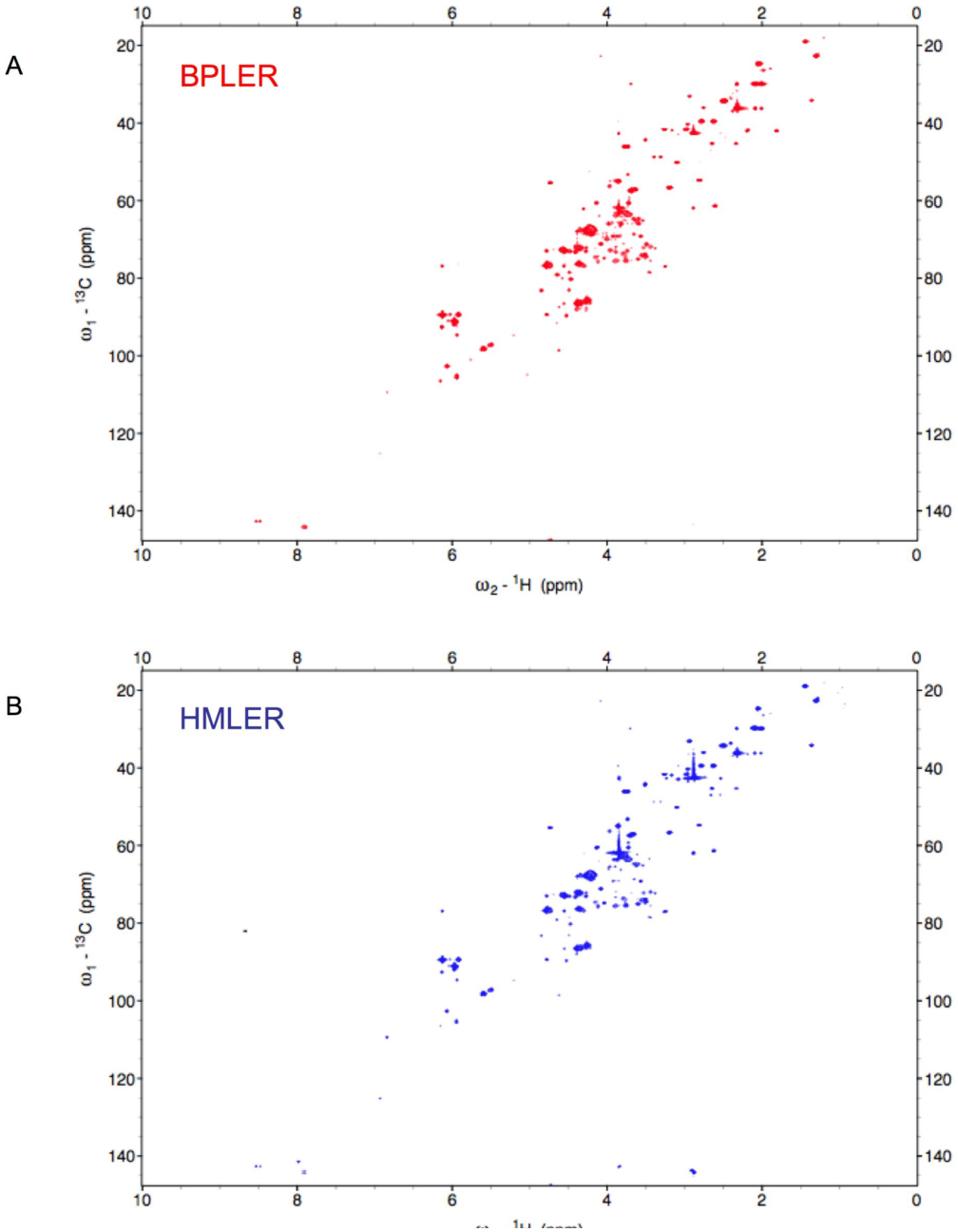


Figure 4.3 ^{13}C-^{1}H HSQC metabolite spectra of BPLER and HMLER cells.

Representative ^{13}C-^{1}H HSQC spectra for ~15 million BPLER (A) and HMLER (B) cells supplemented with ^{13}C-glucose. Metabolite resonance peaks correspond to ^{13}C nuclei attached to unique protons.

derived from the ^{13}C-glucose added to cells (**Figure 4.3**). By overlaying the HSQC spectra from BPLER and HMLER cells we found that the intensity and expression of some metabolite resonances differed between the cell types (**Figure 4.4**). Although many resonances varied in intensity, we decided to focus on the 9 resonances that were unique to BPLER.

N-acetylneuraminic acid is increased in BPLER cells

To identify the metabolite(s) that correspond to the resonances upregulated in BPLER cells, we queried the Human Metabolome database (HMDB). The HMDB contains reference NMR information for more than 7,000 detected metabolites (*8, 9*). For 8 of the 9 ^{13}C-^{1}H frequencies uniquely dected in BPLER cells, known metabolites contain similar resonance frequencies (**Table 4.1**). Six of these resonances are found in N-acetylneuraminic acid (Neu5Ac) (**Figure 4.5**). To verify this assignment the ^{13}C-^{1}H HSQC reference spectrum of pure Neu5Ac was downloaded from the HMBD and compared to the BPLER and HMLER spectra (**Figure 4.6**). Indeed all of the Neu5Ac resonances overlap with resonances specifically increased in the BPLER spectra. This strongly suggests that the differential resonances correspond to Neu5Ac, and that BPLER have increased Neu5Ac. To validate this assignment further, we used liquid chromatography mass spectrometry (LC-M/S). Neu5Ac was used as a standard to determine retention time. BPLER and HMLER water-soluble metabolite extracts were analyzed for Neu5Ac abundance using multiple reaction monitoring (MRM) for two fragments of Neu5Ac. Neu5Ac was ~7-fold more abundant in BPLER cells than in

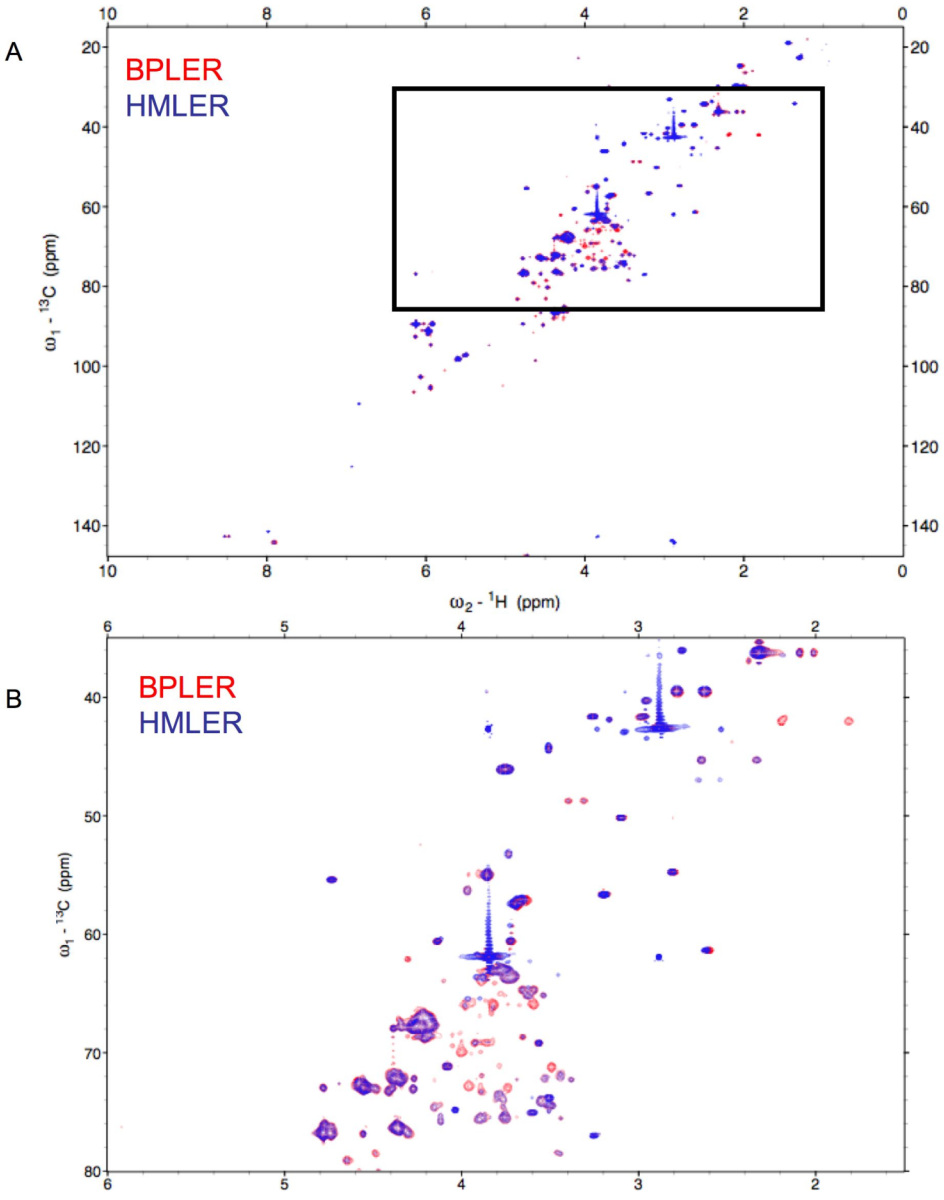


Figure 4.4 BPLER and HMLER show differential production of glucose metabolites. Overlay of ^{13}C-^{1}H spectra for BPLER (red) and HMLER (blue) cells indicated the intensity and expression of metabolite resonances differed between cell types (A). Close up of the boxed region highlights resonances with increased intensity in BPLER cells (red) (B).

Table 4.1 Metabolite resonances enriched in BPLER cells. ^{13}C-^{1}H resonance frequencies detected in BPLER but not in HMLER cells. Potential differentially produced metabolites were identified by searching reference compounds in the Human Metabolome Database with similar resonance frequencies. Chemical shift tolerances of +/- 0.05 ppm and +/- 0.1 ppm were allowed in the ^{1}H and ^{13}C dimension respectively. N-acetylneuraminic acid (red) has resonances at the same frequency as 6 of the 9 BPLER-specific peaks and its precursors N-acetyl-D-mannosamine and uridine 5'-diphospho-N-acetylglucosamine may correspond to another peak (bold).

^{13}C	^{1}H	Potential metabolites
62.993	3.842	**N-acetyl-D-mannosamine**, D-glucosamine, **Uridine 5'-diphospho-N-acetylglucosamine**
65.98	3.823	**N-acetylneuraminic acid**, sucrose, lactulose
68.633	4.248	L-threonine, NAD
69.973	4.002	**N-acetylneuraminic acid**
42.028	2.193	**N-acetylneuraminic acid**, 21-Hydroxypregnenolone
72.951	3.739	**N-acetylneuraminic acid**
72.849	3.953	muco-Inositol, L-arabitol, **N-acetylneuraminic acid**, 2-alpha-Mannobiose, D-pinitol, D-Cellobiose, lactose, Melibiose
42.015	1.814	**N-acetylneuraminic acid**
106.501	6.151	No matches

N-acetylneuraminic acid
(Neu5Ac)

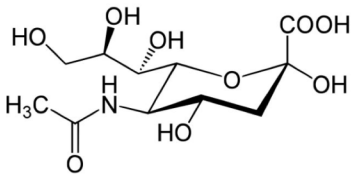


Figure 4.5 Structure of N-acetylneuraminic acid (Neu5Ac). N-acetylneuraminic acid (Neu5Ac) was identified as a putative metabolite that might have many of the differential resonances observed in the BPLER spectra.

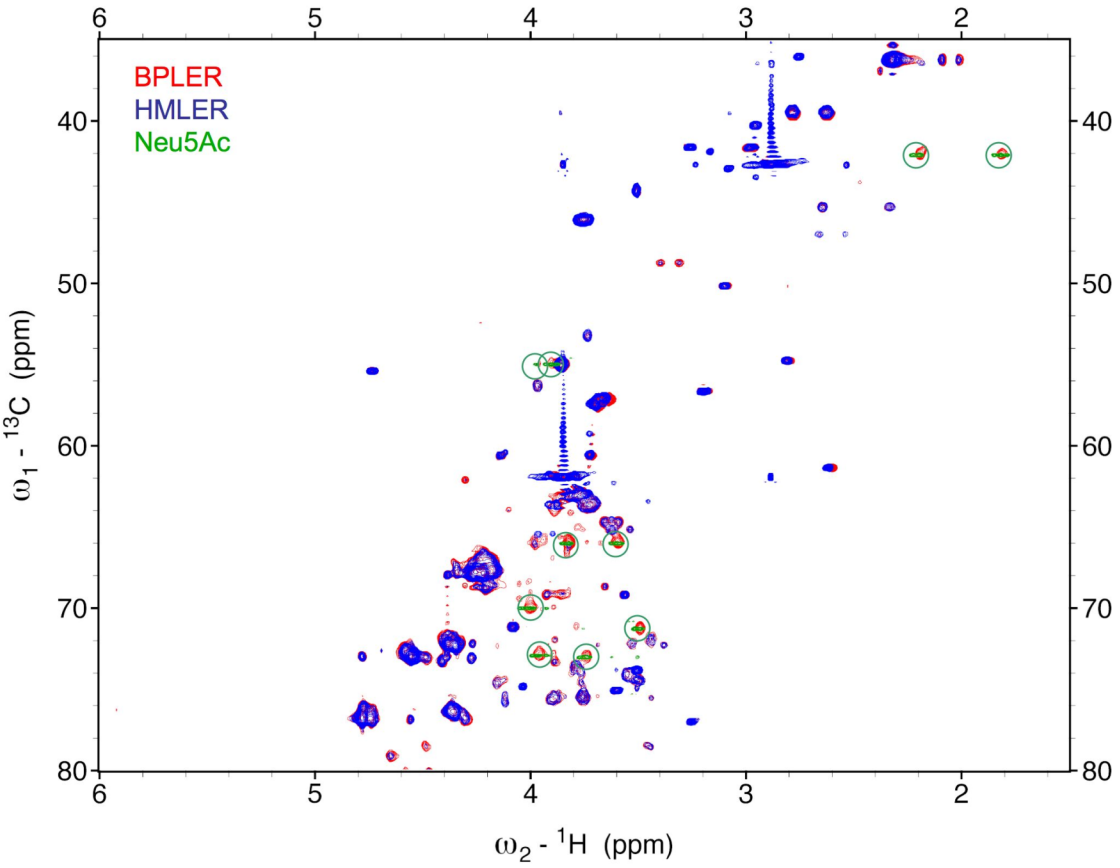


Figure 4.6 Neu5Ac resonances overlap with the differential resonances of BPLER cells. Overlay of a ^{13}C-^{1}H HSQC reference spectrum of pure Neu5Ac (green) with HSQC spectra of BPLER cells (red) and HMLER cells (blue). Neu5Ac resonances that overlap with the differentially expressed resonances of BPLER cells are denoted by green circles.

HMLER cells (**Figure 4.7**). The NMR and M/S results confirmed that Neu5Ac generation is increased in BPLER cells.

BPLER cells have increased Neu5Ac on cell surface glycoproteins

Neu5Ac is the most abundant member of the sialic acid family of sugars (*10*). Sialic acids primarily exist as terminating residues on glycosylation chains of cell surface glycoproteins (*11*). Increased sialic acid levels are associated with metastasis (*12-14*). The mechanism that leads to increased sialic acid expression remains unclear (*15*).

Neu5Ac is generated through the hexosamine pathway, where multiple enzymes convert glucose to UDP-N-acetylglucosamine (UDP-GlcNAc) (*10*). UDP-GlcNAc is the common precursor for all amino sugars of glycoproteins, lipids and proteoglycans. UDP-GlcNAc was also one of the potential metabolites detected only in BPLER cells (**Table 4.1**). UDP-GlcNAc is then converted to Neu5Ac by sequential action of enzymes glucosamine (UDP-N-acetyl)-2-epimerase (GNE), N-acetylneuraminic acid synthase (NANS) and N-acetylneuraminic acid phosphatase (NANP) (*16, 17*). Neu5Ac is then activated by the enzyme cytidine-5-monophosphate N-acetylneuraminic acid synthetase (CMAS) into cytidine-5-monosphosphate N-acetylneuraminic acid (CMP-Neu5Ac), which is transported to the Golgi. In the Golgi sialyltransferases (STs) catalyze the transfer of Neu5Ac onto terminal residues of glycoproteins and glycolipids (**Figure 4.8**).

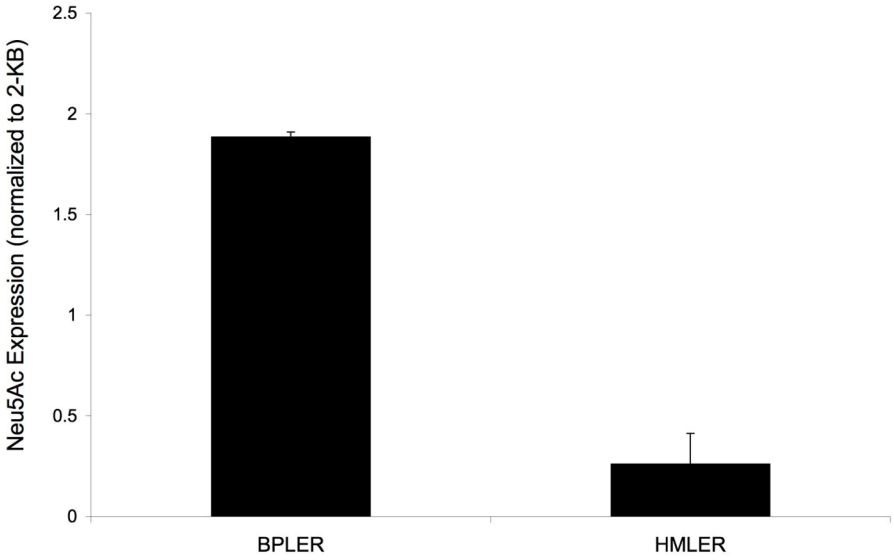


Figure 4.7 Neu5Ac is more abundant in BPLER cells. LC-M/S multiple reaction monitoring (MRM) for two fragments of Neu5Ac demonstrated that Neu5Ac is ~7-fold more abundant in BPLER cells than in HMLER cells. Values are reported as area under the Neu5Ac peak compared to a standard spike-in control of 2-ketobutryic acid (2-KB). Shown are the average values and standard deviation for three replicates. Performed with Colin Hill at Agios Pharmaceutical.

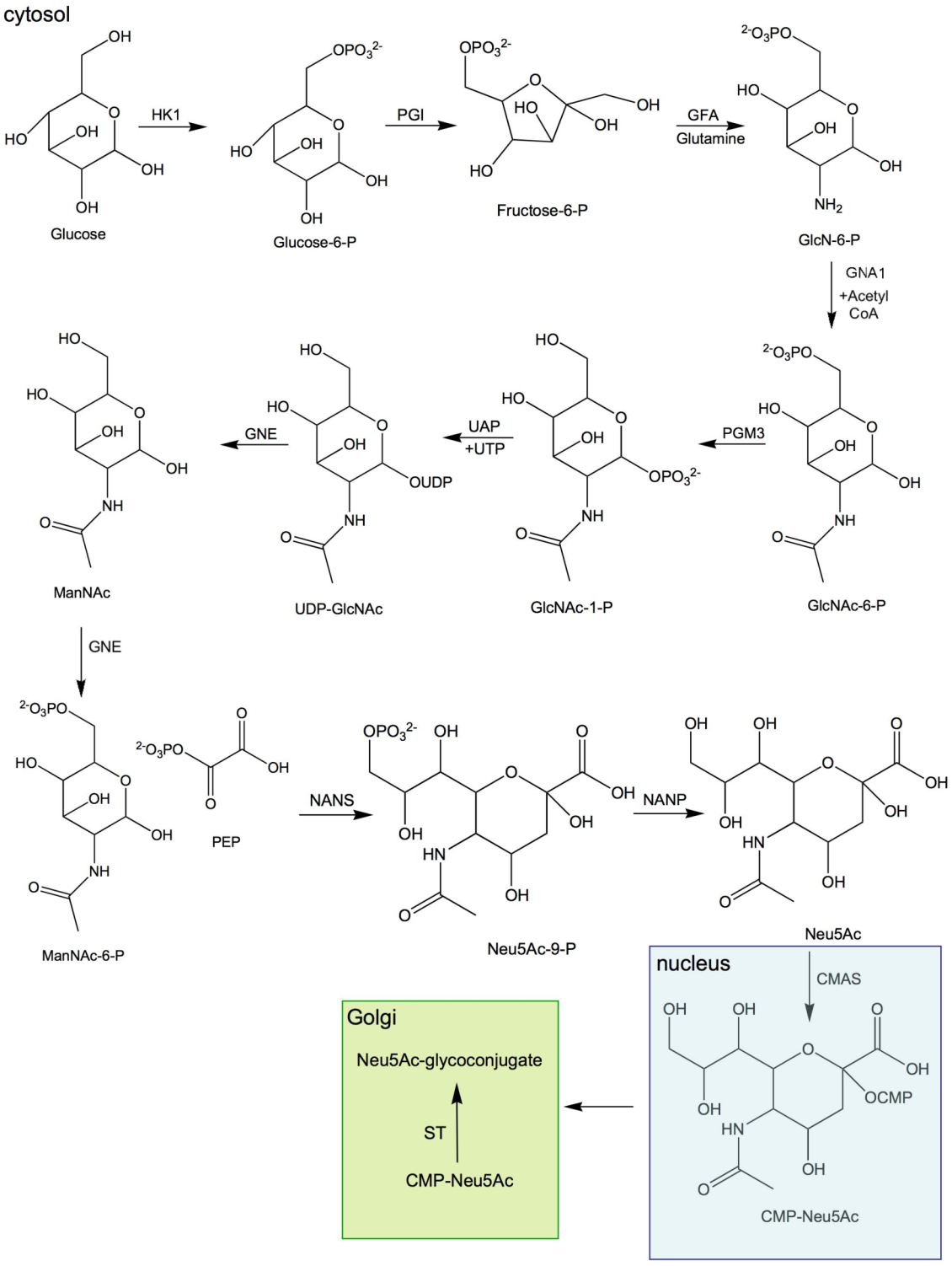


Figure 4.8 Neu5Ac biosynthetic pathway (Continued).

Figure 4.8 Neu5Ac biosynthetic pathway (Continued)**.** In the cytosol glucose is converted to glucose-6-phosphate (glucose-6-P) by hexokinase 1 (HK1). Phosphoglucose isomerase (PGI) converts glucose-6-P into fructose-6-phosphate (fructose-6-P). Utilizing glutamine, glucosamine-fructose-6-phosphate aminotransferase (GFA) next generates glucosamine-6-phosphate (GlcN-6-P). In an acetyl-CoA dependent manner glucosamine-6-P-N-acetyltransferase (GNA1) generates N-acetylglucosamine-6-phosphate (GlcNAc-6-P). Phospho-acetyl-glucosamine mutase (PGM3) produces N-acetylglucosamine-1-phosphate (GlcNAc-1-P). Utilizing UTP, UDP-N-acetylglucosamine pyrophosphorylase (UAP) generates UDP-N-acetylglucosamine (UDP-GlcNAc). In two steps glucosamine (UDP-N-acetyl)-2-epimerase (GNE) converts UDP-GlcNAc to N-acetyl-D-mannosamine (ManNAc) then to N-acetyl-D-mannosamine-6-phosphate (ManNAc-6-P). N-acetylneuraminic acid synthase (NANS) adds phosphoenolpyruvate (PEP) to ManNAc-6-P to generate N-acetylneuraminic acid-9-phosphate (Neu5Ac-9-P). N-acetylneuraminic acid phosphatase (NANP) removes the phosphate to produce N-acetylneuraminic acid (Neu5Ac). In the nucleus, cytidine monophosphate N-acetylneuraminic acid synthase (CMAS) activates Neu5Ac to cytidine monophosphate N-acetylneuraminic acid (CMP-Neu5Ac). In the Golgi, sialyltransferases (STs) utilize CMP-Neu5Ac to attach Neu5Ac onto terminating residues of glycoproteins.

Increased Neu5Ac in BPLER could suggest increased sialylation on BPLER cell surface glycoproteins. To examine this we used wheat-germ-agglutinin (WGA) binding and fluorescence microscopy. WGA binds to Neu5Ac residues on glycoproteins (*18*). Therefore we incubated BPLER and HMLER cells with rhodamine-labeled WGA and compared the amount of rhodamine fluorescence in each cell line (**Figure 4.9**). BPLER cells had increased fluorescence. This may suggest that WGA more readily binds to BPLER cells. Since WGA binds Neu5Ac residues, it is likely BPLER cells are more heavily sialylated.

N-acetylneuraminic acid expression correlates with invasiveness.

Since Neu5Ac has previously been linked to metastasis (*12-14*) we hypothesized that increased Neu5Ac levels might contribute to invasiveness of BPLER cells. To test this hypothesis, we preformed invasion assays. In these assays cells are seeded in an upper chamber and the number of invading cells that penetrate through a matrigel-coated 0.8 μm porous basement membrane into a lower chamber are monitored. Invading cells in the lower chamber are visualized by crystal violet staining and counted. Approximately 5-fold more BPLER cells invaded through the membrane than HMLER cells (**Figure 4.10**). This is consistent with the increased malignancy of BPLER cells (*6*). Neuraminidase specifically removes α2-3, α2-6, and α2-8 linked Neu5Ac residues from cell surface glycoproteins (*10*). Treating BPLER cells with 2 units of neuraminidase completely abolished their ability to invade (**Figure 4.10**). This suggests that Neu5Ac residues on glycoproteins contribute to BPLER invasiveness.

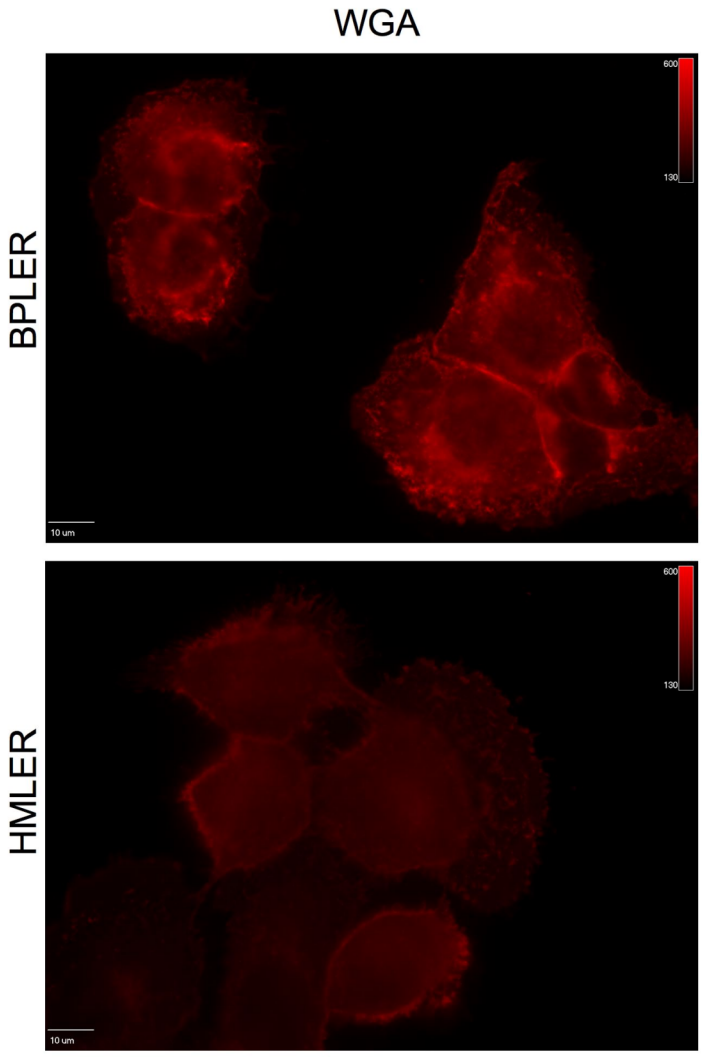


Figure 4.9 Neu5Ac is more abundant on BPLER cell-surface glycoproteins.

BPLER and HMLER cells were grown on coverslips and incubated with rhodamine-labeled wheat-germ-agglutinin (WGA). WGA binds Neu5Ac residues on glycoproteins (18). Rhodamine staining was measured using fluorescence microscopy.

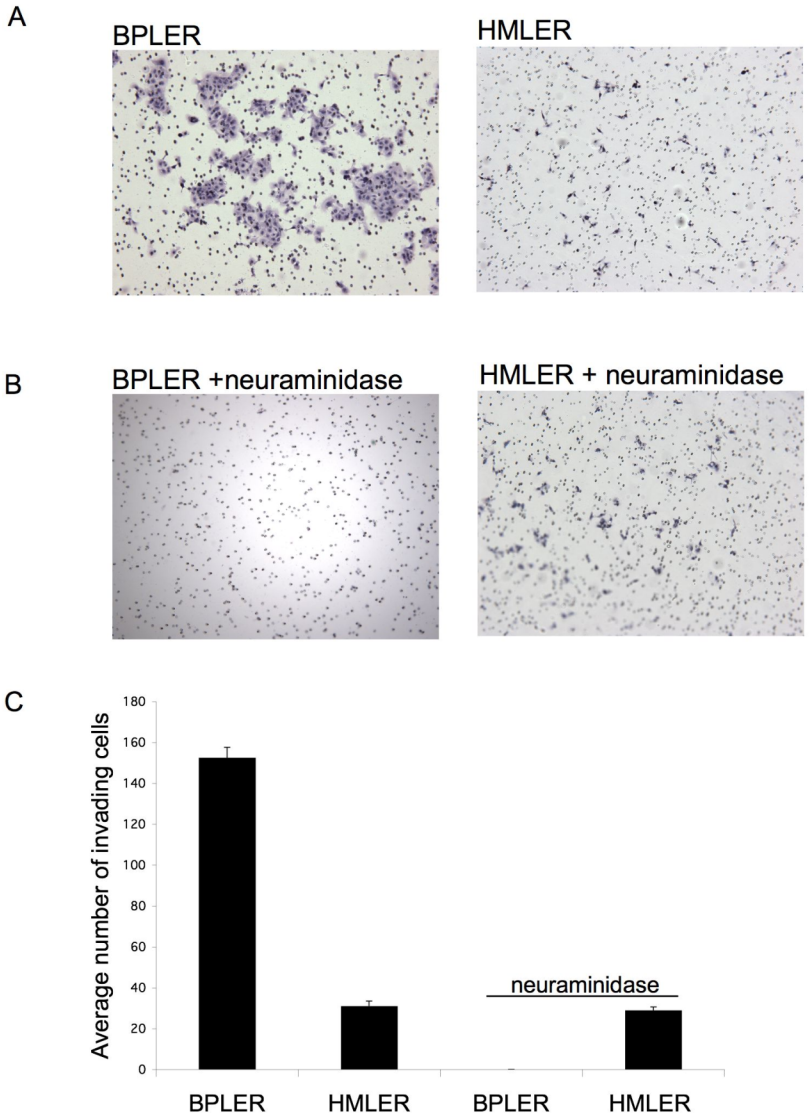


Figure 4.10 Invasiveness of BPLER and HMLER cells correlates with Neu5Ac expression. (A-B) Representative images of the number of invading BPLER and HMLER cells (A) and BPLER and HMLER cells treated with 2 units of neuraminidase (B) that penetrated a 0.8 μm porous basement membrane overlaid with matrigel. Cells were stained with crystal violet. (C) The average number of invading cells and the standard deviation for 3 replicate experiments are shown.

Neuraminidase treatment only marginally decreased the ability of HMLER cells to invade.

CMAS expression is increased in BPLER cells

Increased Neu5Ac in BPLER relative to HMLER could be due to increased Neu5Ac biosynthesis in BPLER cells. To begin to examine this possibility we compared the expression of two of the enzymes required for its biosynthesis, NANS and CMAS. There was no significant difference in NANS or CMAS mRNAs in BPLER compared to HMLER (data not shown). NANS protein levels by western blot also were comparable (**Figure 4.11**). However CMAS protein was readily detected by western blot only in BPLER. CMAS was also readily detected in another aggressive TNBC cell line MB468.

Neu5Ac biosynthetic enzymes, NANS and CMAS, contribute to BPLER cell invasiveness

We next sought to determine if Neu5Ac biosynthetic enzymes NANS and CMAS contribute to the malignant phenotype of BPLER cells. NANS catalyzes the condensation of N-acetylmannosamine (ManNAc-6-P) and phosphoenolpyruvate (PEP) into N-acetylneuraminic acid-9-phosphate (Neu5Ac-9-P) (**Figure 4.8**). ManNAc-6-P was also one of the potential metabolites only detected in BPLER cells (**Table 4.1**). CMAS activates Neu5Ac to CMP-Neu5Ac, which is required for transfer of Neu5Ac onto glycoproteins. To evaluate the contribution of Neu5Ac biosynthesis to invasivity, we transfected BPLER and HMLER with siRNAs against NANS (siNANS) and CMAS

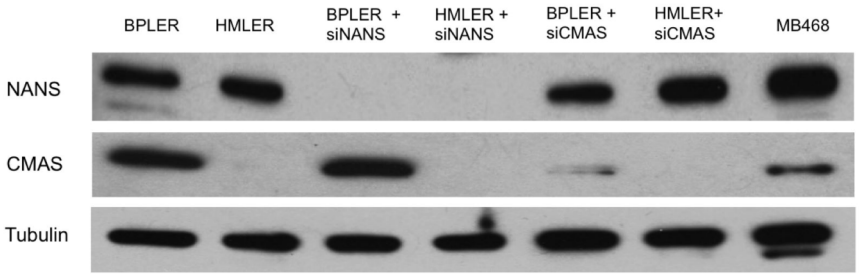


Figure 4.11 CMAS protein is more abundant in BPLER cells. Immunoblots for NANS and CMAS. Knockdown by siRNAs was effective.Tubulin was probed as a loading control.

(siCMAS). Near complete knockdown of both proteins was observed (**Figure 4.11**). Knockdown of either gene had little effect on BPLER proliferation as assessed by counting live cells 3 days after transfection with siNANS, siCMAS or a control siRNA (siControl) (**Figure 4.12**).

When the invasion assays were preformed using BPLER cells treated with siControl, siNANS or siCMAS, knockdown of either CMAS or NANS profoundly inhibited invasion (**Figure 4.13)**. Thus, NANS and CMAS play an important role in promoting BPLER cell invasion. Of note, adding Neu5Ac to the cell medium partially restored invasivity.

CMAS expression correlates with invasion and tumor initiation

To explore further the role of CMAS in invasion, we transfected the less invasive HMLER cells which had undetectable CMAS protein, with a plasmid overexpressing CMAS and repeated the invasion assay. There was a 8-fold increase in invading HMLER cells transfected to overexpress CMAS as compared to wild-type HMLER cells or HMLER cells expressing a control plasmid (**Figure 4.14**). Thus, increased CMAS enhances invasion.

Stable knockdown of CMAS inhibits tumor initiation

We next assessed whether CMAS expression was needed for BPLER tumor initiation. BPLER cells were infected with lentiviral vectors expressing a short hairpin RNA (shRNA) against CMAS (BPLER-shCMAS) or an empty vector (BPLER-EV).

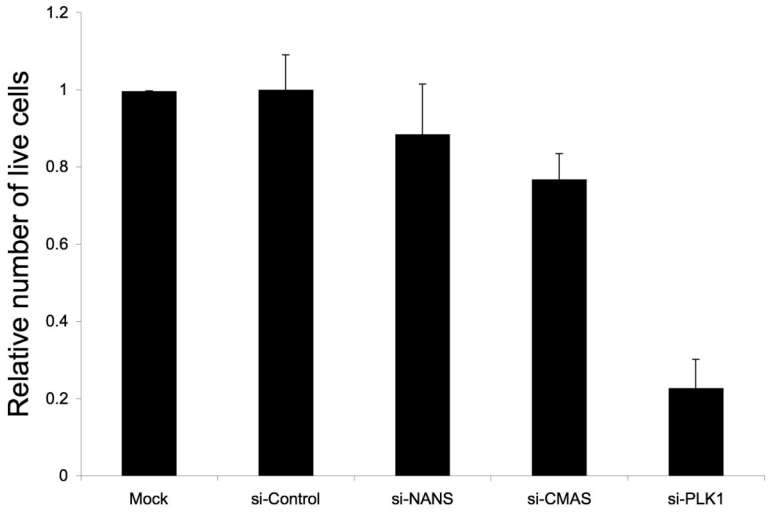


Figure 4.12 siRNA knockdown of NANS and CMAS does not substantially affect BPLER cell viability. BPLER cell numbers were assessed by CellTiterGlo assay 3 days after transfection with a non-targeting siRNA (siControl) or siRNAs against NANS (siNANS) and CMAS (siCMAS). Knockdown of the cell cycle regulator PLK1 (siPLK1) was used as a positive control. Cell numbers were normalized to mock treated cells. Shown are average and standard deviation for 3 replicate experiments.

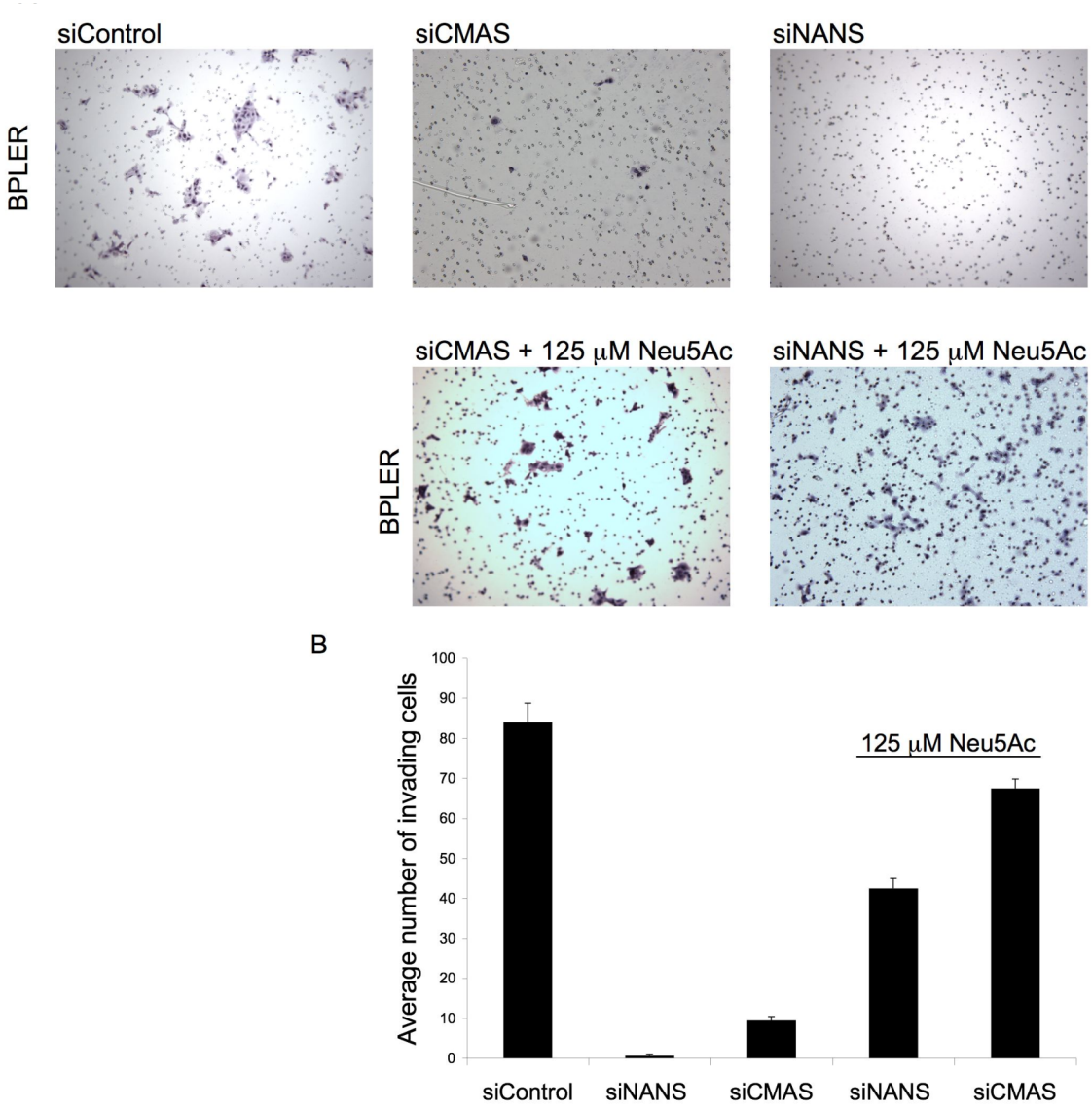


Figure 4.13 NANS and CMAS expression are needed for BPLER invasiveness. (A) Representative images of invading BPLER cells after transfection with control siRNA (siControl) or siRNAs against NANS (siNANS) or CMAS (siCMAS). In some cells 125 μM was Neu5Ac added to the medium. (B) Average number of invading cells in each condition and standard deviation for 3 experimental replicates are shown.

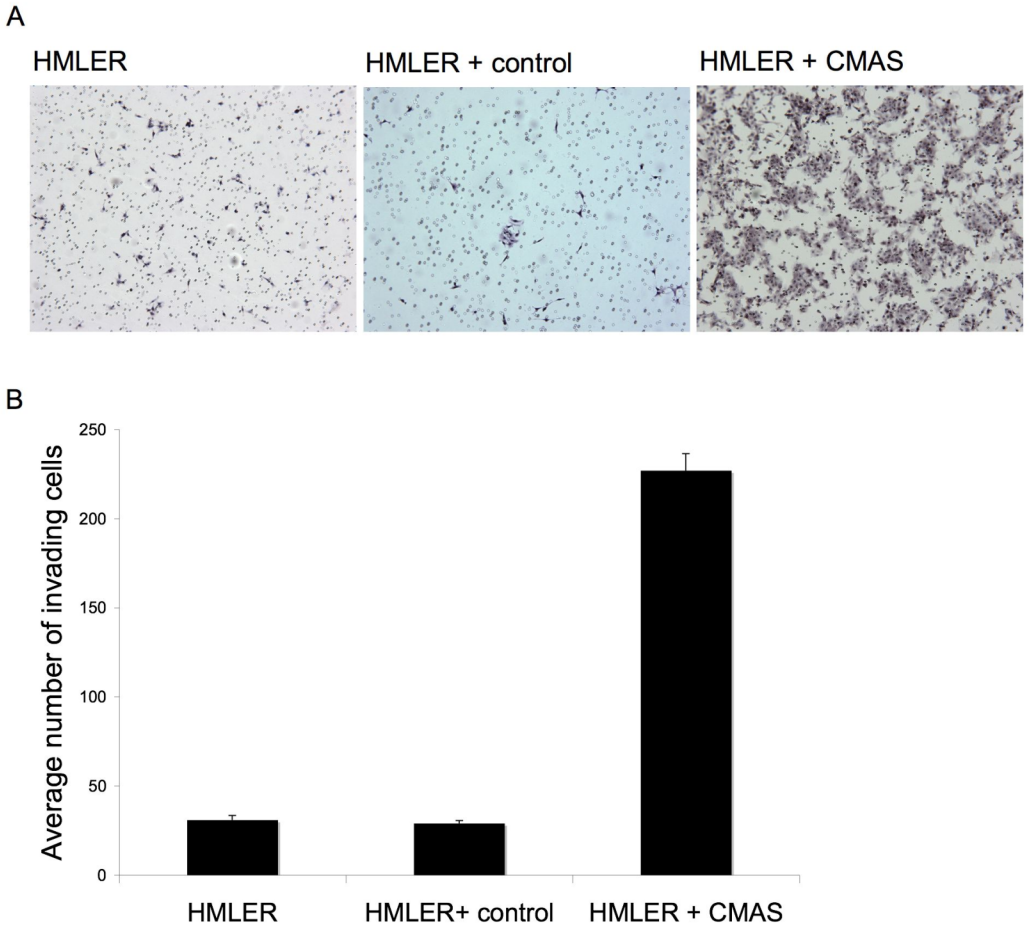


Figure 4.14 Overexpressing CMAS in HMLER cells increases migration. (A) Representative images of the number of invading HMLER cells and HMLER cells transfected with a control plasmid or a plasmid overexpressing CMAS that penetrate the matrigel coated 0.8 μm porous basement membrane. Cells were stained with crystal violet. (B) The average number of migrating cells and standard deviation of three replicate experiments are shown.

The lentiviral vectors contained a RFP selection marker, which was used to sort cells that contained the integrated vector. BPLER-shCMAS cells had 90% less CMAS than BPLER-EV cells by qRT-PCR (**Figure 4.15A**). No CMAS protein was detected in BPLER-shCMAS cells by immunoblot (**Figure 4.15B**).

$5x10^5$ BPLER-shCMAS or BPLER-EV cells were injected subcutaneously into the flanks of 4-week old NOD/SCID/IL2R$\gamma^{-/-}$ mice (N=4/group). Every 3 days tumor size was monitored (**Figure 4.16**). After ~60 days, the BPLER-EV tumors exceeded 15 mm in diameter and the mice were sacrificed. During this same time period, no palpable tumors were detected in mice injected with BPLER-shCMAS cells. This suggests that CMAS expression is critical for tumorigenesis.

Discussion

The metabolic output of a cell is influenced by the functional genomic, transcriptomic and proteomic networks that define that cell type. Several analytical techniques are available to profile metabolites including mass-spectrometry (M/S), chromatography and NMR spectroscopy. From small amounts of starting material M/S and chromatography can provide absolute quantification of metabolites. However in most cases the metabolites of interest must be selected *a priori*. Untargeted M/S approaches typically involve multiple rounds of purification and identification steps. NMR has the advantage that it can provide global and unbiased metabolite structural information. However, NMR is much less sensitive and generally requires increased

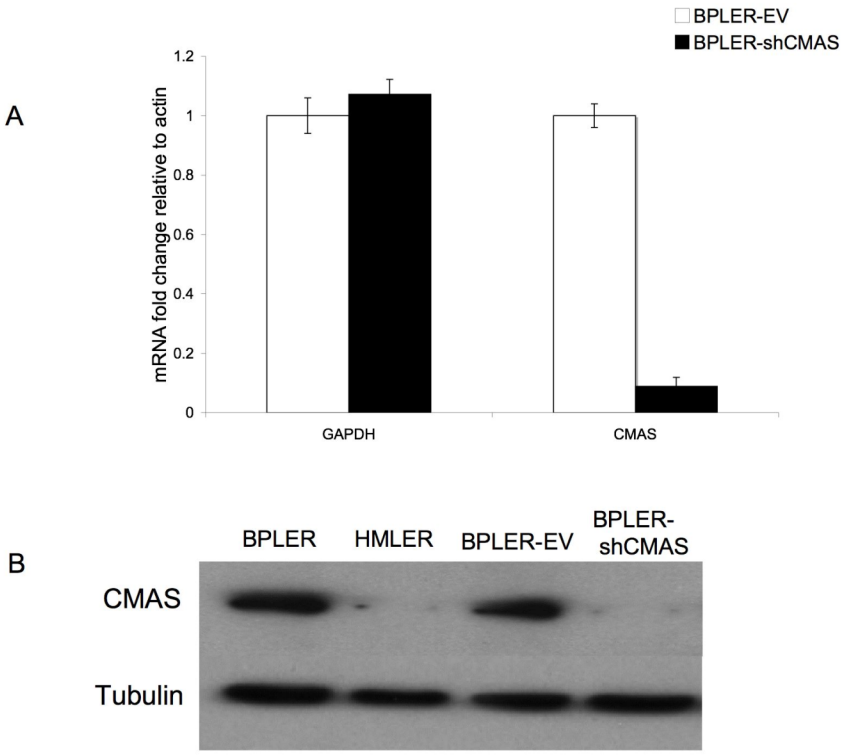


Figure 4.15 BPLER-shCMAS cells have decreased CMAS expression. (A). qRT-PCR and (B) western blot to compare CMAS mRNA and protein respectively. BPLER-shCMAS cells had decreased CMAS expression.

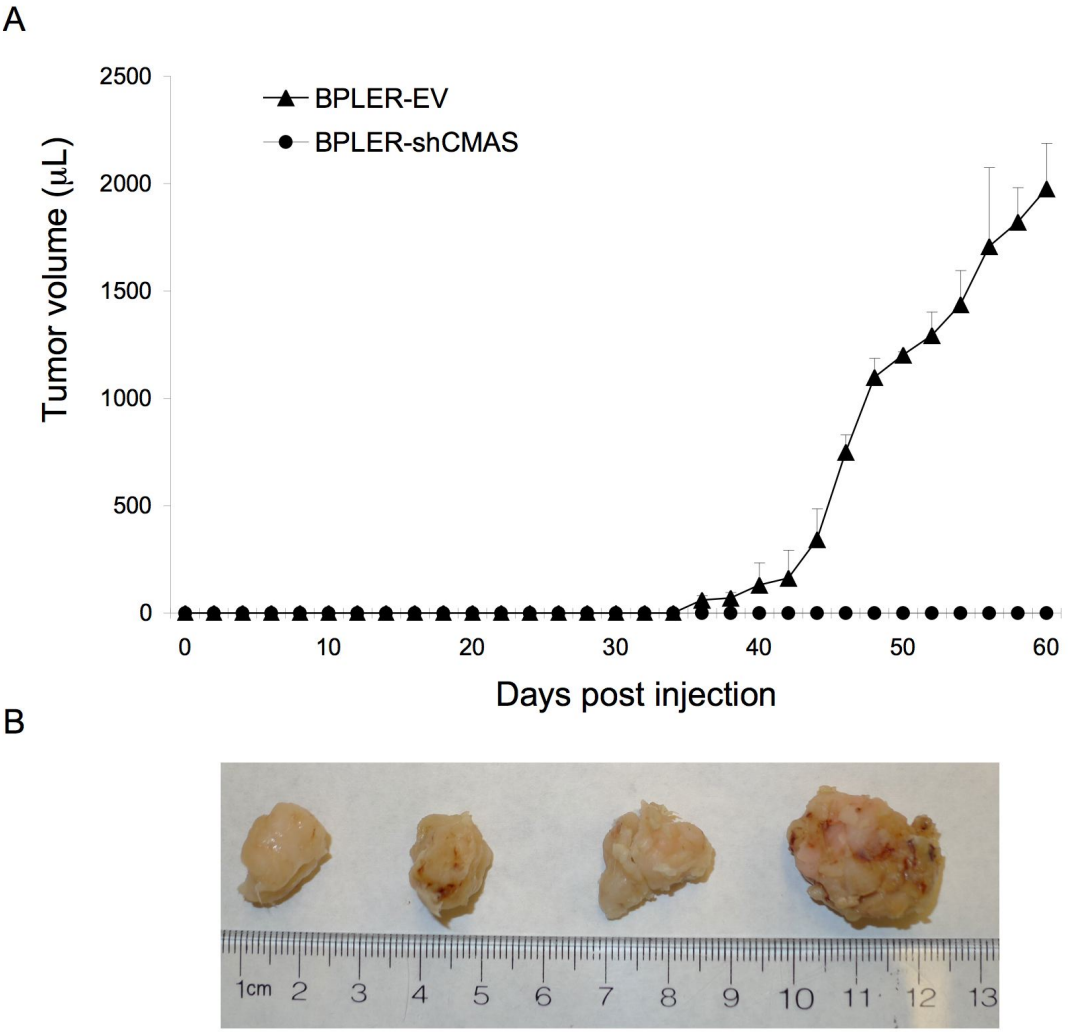


Figure 4.16 CMAS expression is required for tumor initiation. $5x10^5$ BPLER-EV or BPLER-shCMAS cells were injected subcutaneously into the flanks of 4-week old NOD/SCID/IL2Rγ^{null} mice (N=4/group). (A) tumor size was recorded every 3 days. The average tumor volume and standard deviation for each group of mice are shown. (B) Tumors from BPLER-EV injected mice exceeded 15 mm in diameter by day 60 and the mice were sacrificed. The tumors are shown. No palpable tumors were detected in mice injected with BPLER-shCMAS cells during this time period.

sample amounts. This can be incredibly burdensome, especially for 2D NMR experiments that rely on 1.1% ^{13}C natural abundance.

Here we utilized an unbiased NMR metabolite-screening platform to specifically follow glucose catabolism in the highly tumorigenic BPLER cell line and compare it to glucose products in the less aggressive HMLER cell line. Instead of using large sample amounts we used relatively few cells (~15 million) fed with ^{13}C-glucose. With these conditions, background ^{13}C was negligible and we could follow the products of the ^{13}C-labeled tracer. We recorded ^{13}C-^{1}H HSQC spectra to uncover the glucose metabolites in BPLER and HMLER cell lines. By comparing spectra we observed that most resonances agreed but a small number of metabolite resonances varied between the cell lines suggesting that BPLER and HMLER cells differentially catabolize glucose.

Nine metabolite resonances were only detected in BPLER cells, a similar number were unique to HMLER cells. Some resonances were detected in both BPLER and HMLER cells but with different intensities. To gain a complete understanding of the metabolic pathways in these cells, future work could examine all of these differences. However, here we focused on the metabolite resonances unique to BPLER cells. We matched these resonance with reference compounds in the Human Metabolome Database (HMBD), which identified N-acetylneuraminic acid (Neu5Ac) as a potential metabolite enriched in BPLER cells. All the resonances in the reference ^{13}C-^{1}H HSQC spectra of Neu5Ac corresponded to resonances only detected in BPLER cells. LC-M/S

analysis further confirmed that Neu5Ac is ~7-fold more abundant in BPLER cells as compared to HMLER cells.

Neu5Ac is the most abundant sialic acid family sugar that is added as the terminal residue to some cell surface glycoproteins (*10, 11, 19, 20*). Neu5Ac expression confers mucinous properties to epithelial cell surfaces, providing a lubricant that makes them “slimy” and interferes with adhesion (*14, 21, 22*). Perhaps it should not be surprising that cells with more Neu5Ac have a more sialylated surface and be more invasive. In fact we found that BPLER were much more heavily sialyated than HMLER by staining with wheat germ agglutinin. Sialylation mediates numerous biological processes including masking antigenic determinates (*23, 24*), and altering cell-cell recognition, cell-contact inhibition (*25*), and cellular migration (*10, 26*). Increased Neu5Ac has been observed in many metastatic tumors (*12-14, 27*). Consistent with the literature, we observed Neu5Ac expression correlated with cellular invasiveness. Using invasion assays we showed that more BPLER cells invaded through a 0.8 μm porous membrane overlaid with matrigel as compared to HMLER cells. The ability of highly malignant BPLER cells to invade through the membrane was lost after neuraminidase treatment. Neuraminidase treatment only slightly reduced the ability of HMLER cells to invade. HMLER cells produce much less Neu5Ac and have much less surface sialylation. Identifying BPLER-specific Neu5Ac containing proteins in the future may pinpoint specific glycoproteins that are critical for tumor initiation and metastasis.

siRNA knockdown of the Neu5Ac biosynthetic enzymes NANS and CMAS drastically reduced the number of invading BPLER cells. Addition of Neu5Ac to the cell

medium partially restored invasion. The mechanism by which Neu5Ac rescues migration is unclear. One possibility is that sialyltransferases in serum catalyzed the addition of Neu5Ac onto cell surface proteins.

CMAS protein expression was markedly increased in 2 highly tumorigenic BPLER and MB468 epithelial TNBC cell lines. Because TNBC cell lines and tumors are very heterogeneous, we are currently assessing CMAS mRNA and protein expression to determine if this is a common feature of epithelial TNBC cell lines. Overexpressing CMAS in HMLER cells increased their invasiveness. Reciprocally BPLER cells with stably knocked down CMAS were unable to form tumors in mice. These results suggest that CMAS is a critical regulator of invasion and tumor initiation in BPLER. It will be worthwhile to determine if CMAS expression is required for tumor formation and invasion for other TNBCs. We also plan to investigate in the future whether CMAS overexpression enhances tumor formation by HMLER cells.

CMAS is expressed in both eukaryotic and prokaryotic organisms (*28*). CMAS catalyzes the conversion of Neu5Ac to CMP-Neu5Ac. This activation step is required to transfer Neu5Ac to glycoproteins. In pathogenic bacteria CMAS is a virulence factor catalyzing the production of sialylated proteins that mimic host proteins (*28, 29*). In humans, a prominent role of CMAS in cancer development has never before been noted. A single report, using cancer-specific high-throughput annotation of somatic mutations (CHASM), identified genetic mutations in 18 genes including CMAS important for pancreatic cancer (*30*). No functional studies were performed. Our findings suggest that CMAS expression is essential for tumor initiation. Thus CMAS could be a novel

therapeutic target for some breast cancers. The crystal structure of the catalytic domain of murine CMAS has been resolved to 2.8 Å (*31*). This could serve as the starting point for rational drug design of CMAS inhibitors.

Summary of this work

In part 2 of this thesis we used an unbiased NMR metabolite screening method to identify glucose metabolites differentially produced in BPLER and HMLER isogenic triple negative breast cancer cell lines that have dramatic differences in tumor initiating capacity. N-acetylneuraminic acid (Neu5Ac) a terminating sugar of many glycoproteins is much more abundant in BPLER than in HMLER cells. Manipulating Neu5Ac expression by enzymatic removal or knockdown of biosynthetic enzyme N-acetylneuraminic acid synthase (NANS) and cytidine monophosphate N-acetyneuraminic acid synthetase (CMAS) reduces *in vitro* invasivity of BPLER cells. CMAS is increased in BPLER cells. Overexpressing CMAS in HMLER cells increases their invasiveness. Stable knockdown of CMAS blocks BPLER tumor growth in mice. From our results we demonstrated increased Neu5Ac and CMAS expression is linked to tumor initiation and invasivity in a human triple negative breast cancer cell line.

Materials & Methods:

Cell culture

BPLER and HMLER cells were provided by R. Weinberg and Tan Ince and grown in WIT medium (Stemgent). MB468 cells were obtained from ATCC and grown in DMEM supplemented with 10% FBS. For all transfections BPLER cells were reverse-transfected with 50 nM siRNA using Dharmafect#1 (Dharmacon) in WIT medium. siGenome Smartpool siRNAs for human NANS (M-013399) and CMAS (M-009780) and Non-Targeting siRNA #1 (D-001210) were obtained from Thermo Scientific (Waltham, MA).

NMR sample preparation

For labeling experiments, $1.5x10^7$ BPLER and HMLER cells were grown in WIT medium. Cells were washed twice with warm PBS and then incubated in OPTIMEM medium without glucose for ~30 min. Unlabeled or uniformly labeled ^{13}C-glucose (Cambridge Isotope Laboratories, Cambridge, MA) was added directly to the medium and cells were incubated an additional ~2 hrs. Cells were wash twice with PBS, trypsinized with Tryple Express (Invitrogen), and counted using Trypan blue staining. Equal numbers of cells were harvested by centrifugation at 500 x g for 5 min. Cell pellets were immediately flash frozen in liquid nitrogen and stored at -80°C until ready for analysis. Cell pellets were thawed on ice and lysed by the addition of ice cold methanol and vortexed vigorously. Equal parts chloroform and water were added to the extract and vortexed vigorously. Extracts were incubated overnight at -20°C for phase

separation. The extract mixture was centrifuged at 10,000 x g for 40 min to complete phase separation and separate macromolecules. After centrifugation three layers were visible. The bottom phase contained organic metabolites. The interphase layer (white clumps) contained proteins and macromolecules. The top 2/3 water:methanol layer contained water soluble metabolites. The top layer was collected and methanol was removed under vacuum. The water metabolites were flash frozen with liquid nitrogen and lyophilized. The completely dried samples were resuspended in 250 μL D_2O containing 1 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) standard (Cambridge Isotope Laboratories). Samples were run in Shigemi tubes (Bruker).

NMR spectroscopy

The NMR spectra were collected on a Bruker 500 MHz spectrometer equipped with a cryoprobe. 2D ^{13}C-^{1}H HSQC spectra were collected with a relaxation delay of 1 s. A total of 2048 points with a spectral width of 4734.9 Hz and 13834.3 Hz were collected in the ^{1}H and ^{13}C dimensions respectively with 128 data points. NMR data were processed using NMRPipe and Sparky (T.D. Goddard and D.G. Kneller, SPARKY3, University of California, San Francisco) to identify chemical shift resonances. Spectra were aligned and intensities scaled using resonances from the 1 mM DSS control.

LC-Mass spectrometry

Standards of N-acetylneuraminic acid (Neu5Ac) were used to optimize chromatography and determine retention time using a Waters Acquit UPLC coupled to an AbSciex-4000

mass spectrometer operating in the negative mode. Reverse phase chromatography was performed using a Waters Acquity HSS T3 2.1x50 mm column with 1.8 micron particle size with mobile phase A of 15 mM acetic acid, 10 mM tributylamine, 5% methanol and mobile phase B of 100% methanol. Neu5Ac was detected using multiple reaction monitoring (MRM) for fragments 307.987 and 307.9170. 2-ketobutyric acid (1 μg/mL) was used as an injection standard. $3x10^5$ BPLER and HMLER water-soluble metabolites were prepared as for the NMR samples. Dried metabolites were dissolved in 1 mL HPLC-grade water. Neu5Ac levels were measured using 10 μL injections.

Fluorescence microscopy

10,000 BPLER and HMLER cells were grown on glass cover slips, washed with warm PBS and fixed with 4% paraformaldehyde (PFA) for 20 min at room temperature. Cells were washed twice and incubated with rhodamine-labeled wheat germ aggultinin (WGA) (Vector Laboratories, Burlingame, CA) in PBS at a 1:1000 dilution for 10 mins. Cells were washed 5 times with PBS then mounted onto slides. Images were acquired using a Zeiss Axiovert 200M fluorescent microscope.

Invasion Assays

Cells were trypsinized and added ($1.25x10^5$ cells/well) in WIT medium to three wells of BD BioCoat™ Matrigel Invasion Chambers (BD Bioscience). WIT medium containing 10% FCS was added to the lower chamber. The invasion chambers were processed 24 hrs later as per the manufacturer's protocols. Invading cells were stained with crystal

violet. Five random fields from each of the triplicate invasion assays were counted using phase contrast microscopy.

Neuraminidase treatment

As for the invasion assays, cells were trypsinized and added ($1.25x10^5$ cells/well) in WIT medium to three wells of BD BioCoat™ Matrigel Invasion Chambers (BD Bioscience) with 2 units of neuraminidase (New England Biolabs). The number of invading cells were stained with crystal violet and processed as for the invasion assays.

Cell viability

Cell viability was assessed using a CellTiterGlo kit (Promega) according to the manufacturer's protocol. Chemiluminescence was measured using a BioTek Synergy2 plate reader.

RNA analysis

Total RNA was extracted with Trizol (Invitrogen) and cDNA prepared from 1 μg total RNA using Thermoscript RT kit (Invitrogen) as per the manufacturer's instructions. 2.5 μl of diluted cDNA (1:20) was used as template for qPCR using Power Sybr-Green Master Mix (Applied Biosystems) and Biorad C1000 Thermal Cycler (Biorad). Relative CT values were normalized to β-actin and converted to a linear scale using the $-\Delta CT$ method.

Western blot

Protein lysates were prepared using RIPA buffer (150 mM NaCl, 1.0% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris pH 8.0, containing Complete Mini-protease Inhibitor Cocktail (Roche)). Protein concentration was determined using the BioRad DC protein assay kit (BioRad) and samples were resolved on 10% SDS-PAGE gels and transferred using a Transblot semi-dry transfer apparatus (BioRad). Blots were probed with antibodies to NANS (Abcam mouse pAb ab88899), CMAS (Abcam mouse mAb ab57454) and α-Tubulin (Sigma mouse mAb clone B-5-1-2). Antibodies were diluted in 5% milk in TBS-T and incubated overnight at 4°C. Secondary mouse HRP-conjugated antibodies were from Amersham. Protein signal was detected using the ECL Plus kit (Amersham).

Stably knocked down cell lines

To generate cells with stably knocked down CMAS we used a modified pGreenPuro lentiviral vector (System Biosystems, Mountain, CA) in which GFP was replaced with RFP (pRFPpuro) vector. Oligos containing the CMAS siRNA sequence (5'-GTGTATGGGTTTCGACAGA-3') and the complementary sequence (anti-sense CMAS) were synthesized into a short-hairpin RNA as (GATCC-siCMAS-CTTCCTGTCAGA-anti-senseCMAS-TTTTTG). Oligos were annealed and ligated into pRFPpuro. Resulting plasmids were verified by sequencing and used to transfect HEK293T packaging cells to generate lentivirus. Expression of RFP were used to sort cells by flow cytometry (FACSAria, Beckton Dickinson). After two rounds of sorting, we obtained a nearly

homogenous cell line stably knocked down for CMAS. RNA and protein expression were reduced by >90% as verified by qRT-PCR and immunoblot.

Tumor models

All animal procedures were performed with IACUC approval. Exponentially growing cells were trypsinized with Tryple Express (Invitrogen) and $5x10^5$ cells were resuspended in a 1:1 WIT-matrigel solution and injected subcutaneously in the flanks of 4-week old NOD/SCID/IL2R$\gamma^{-/-}$ mice. Mice were palpated for tumor formation and tumor size was measured every 3 days by caliper. Tumor volumes were calculated as Volume (μL)=L x W^2 X 0.4. Mice were sacrificed when the tumor size exceeded 15 mm in diameter.

References:

1. Vander Heiden, M. G., Cantley, L. C., and Thompson, C. B. (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation, *Science 324*, 1029-1033.

2. Vander Heiden, M. G., Lunt, S. Y., Dayton, T. L., Fiske, B. P., Israelsen, W. J., Mattaini, K. R., Vokes, N. I., Stephanopoulos, G., Cantley, L. C., Metallo, C. M., and Locasale, J. W. (2011) Metabolic pathway alterations that support cell proliferation, *Cold Spring Harb Symp Quant Biol 76*, 325-334.

3. Lehninger, A. L., Nelson, D. L., and Cox, M. M. (1993) *Principles of Biochemistry*, Worth, New York.

4. Warburg, O. (1956) On the origin of cancer cells, *Science 123*, 309-314.

5. Vander Heiden, M. G., Locasale, J. W., Swanson, K. D., Sharfi, H., Heffron, G. J., Amador-Noguez, D., Christofk, H. R., Wagner, G., Rabinowitz, J. D., Asara, J. M., and Cantley, L. C. (2010) Evidence for an alternative glycolytic pathway in rapidly proliferating cells, *Science 329*, 1492-1499.

6. Ince, T. A., Richardson, A. L., Bell, G. W., Saitoh, M., Godar, S., Karnoub, A. E., Iglehart, J. D., and Weinberg, R. A. (2007) Transformation of different human breast epithelial cell types leads to distinct tumor phenotypes, *Cancer Cell 12*, 160-170.

7. Cavanagh, J., Fairbrother, W. J., Palmer, A. G., and Skelton, N. J. (1996) Protein NMR Spectroscopy Principles and Practice.

8. Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009) HMDB: a knowledgebase for the human metabolome, *Nucleic Acids Res 37*, D603-610.

9. Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J., and Querengesser, L. (2007) HMDB: the Human Metabolome Database, *Nucleic Acids Res 35*, D521-526.

10. Varki, A., Cummngs, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E. (2009) Essentials of Glycobiology.

11. Schauer, R. (2009) Sialic acids as regulators of molecular and cellular interactions, *Curr Opin Struct Biol 19*, 507-514.

12. Lin, S., Kemmner, W., Grigull, S., and Schlag, P. M. (2002) Cell surface alpha 2,6 sialylation affects adhesion of breast carcinoma cells, *Exp Cell Res 276*, 101-110.

13. Bolscher, J. G., Schallier, D. C., van Rooy, H., Storme, G. A., and Smets, L. A. (1988) Modification of cell surface carbohydrates and invasive behavior by an alkyl lysophospholipid, *Cancer Res 48*, 977-982.

14. Dennis, J., Waller, C., Timpl, R., and Schirrmacher, V. (1982) Surface sialic acid reduces attachment of metastatic tumour cells to collagen type IV and fibronectin, *Nature 300*, 274-276.

15. Harduin-Lepers, A., Krzewinski-Recchi, M.-A., Colomb, F., Foulquier, F., Groux-Degroote, S., and Delannoy, P. (2012) Sialyltransferases functions in cancers, *Front Biosci (Elite Ed) 4*, 499-515.

16. Kean, E. L., and Roseman, S. (1966) The sialic acids. X. Purification and properties of cytidine 5'-monophosphosialic acid synthetase, *J Biol Chem 241*, 5643-5650.

17. Blacklow, R. S., and Warren, L. (1962) Biosynthesis of sialic acids by Neisseria meningitidis, *J Biol Chem 237*, 3520-3526.

18. Chazotte, B. (2011) Labeling membrane glycoproteins or glycolipids with fluorescent wheat germ agglutinin, *Cold Spring Harb Protoc 2011*, pdb.prot5623.

19. Audry, M., Jeanneau, C., Imberty, A., Harduin-Lepers, A., Delannoy, P., and Breton, C. (2011) Current trends in the structure-activity relationships of sialyltransferases, *Glycobiology 21*, 716-726.

20. Vann, W. F., Tavarez, J. J., Crowley, J., Vimr, E., and Silver, R. P. (1997) Purification and characterization of the Escherichia coli K1 neuB gene product N-acetylneuraminic acid synthetase, *Glycobiology 7*, 697-701.

21. Brockhausen, I., Yang, J. M., Burchell, J., Whitehouse, C., and Taylor-Papadimitriou, J. (1995) Mechanisms underlying aberrant glycosylation of MUC1 mucin in breast cancer cells, *Eur J Biochem 233*, 607-617.

22. Higa, H. H., and Paulson, J. C. (1985) Sialylation of glycoprotein oligosaccharides with N-acetyl-, N-glycolyl-, and N-O-diacetylneuraminic acids, *J Biol Chem 260*, 8838-8849.

23. Schauer, R. (1982) Chemistry, metabolism, and biological functions of sialic acids, *Adv Carbohydr Chem Biochem 40*, 131-234.

24. Ashwell, G., and Morell, A. G. (1974) The role of surface carbohydrates in the hepatic recognition and transport of circulating glycoproteins, *Adv Enzymol Relat Areas Mol Biol 41*, 99-128.

25. Revilla-Nuin, B., Reglero, A., Feo, J. C., Rodriguez-Aparicio, L. B., and Ferrero, M. A. (1998) Identification, expression and tissue distribution of cytidine 5'-monophosphate N-acetylneuraminic acid synthetase activity in the rat, *Glycoconj J 15*, 233-241.

26. Alhadeff, J. A., and Holzinger, R. T. (1982) Sialyltransferase, sialic acid and sialoglycoconjugates in metastatic tumor and human liver tissue, *Int J Biochem 14*, 119-126.

27. Yogeeswaran, G., and Salk, P. L. (1981) Metastatic potential is positively correlated with cell surface sialylation of cultured murine tumor cell lines, *Science 212*, 1514-1516.

28. Mizanur, R. M., and Pohl, N. L. (2008) Bacterial CMP-sialic acid synthetases: production, properties, and applications, *Appl Microbiol Biotechnol 80*, 757-765.

29. Hao, J., Balagurumoorthy, P., Sarilla, S., and Sundaramoorthy, M. (2005) Cloning, expression, and characterization of sialic acid synthases, *Biochem Biophys Res Commun 338*, 1507-1514.

30. Carter, H., Samayoa, J., Hruban, R. H., and Karchin, R. (2010) Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM), *Cancer Biol Ther 10*, 582-587.

31. Krapp, S., Münster-Kühnel, A. K., Kaiser, J. T., Huber, R., Tiralongo, J., Gerardy-Schahn, R., and Jacob, U. (2003) The crystal structure of murine CMP-5-N-acetylneuraminic acid synthetase, *J Mol Biol 334*, 625-637.

Appendix: Other research contributions

NMR-BASED METABOLITE SCREENING PLATFORM

STATEMENT AS TO FEDERALLY SPONSORED RESEARCH

The inventions were made with Government support under R21 AI087431 awarded by the National Institutes of Health. The government has certain rights in the inventions.

FIELD OF THE INVENTION

The invention relates to NMR-based screening platforms.

BACKGROUND OF THE INVENTION

The metabolic output of a cell is the summation of the functional genomic, transcriptomic and proteomic networks that define that cell type. Metabolomics is the comprehensive and simultaneous systematic determination of metabolite levels in the metabolome and their changes over time as a consequence of stimuli. While other fields may provide information, for example, regarding the copy number of a given gene, mRNA or protein; this study of chemical processes involving metabolites provides the downstream summation of all aberrant genes, RNAs, and/or proteins. This 'metabolic fingerprint' represents a snapshot of all the functioning or non-functioning pathways in a particular cell type.

Several analytical methods including mass spectrometry, chromatography, and NMR spectroscopy have been used to quantify cellular metabolites. Mass spectrometry and chromatography both require small sample amounts and can be easily adapted for high throughput analysis; however, both methods typically involve at least one if not several purification steps. Furthermore, in most cases the metabolites to be examined must be pre-selected *a priori*. Untargeted mass spectrometry approaches are possible but require several rounds of purification and further identification methods. In addition, not all metabolites, including nucleotide analogs and lipids, are easily ionizable and thus cannot be detected via mass spectrometry. Further, the fragmentation pattern resulting from mass spectrometry is not always suitable to distinguish between molecules such as sugars that have equal mass, but different structures, hence limiting the analysis.

SUMMARY OF THE INVENTION

Described herein is a rapid, unbiased, ultra-high resolution, quantitative NMR screening platform that utilizes any one, two, three, four, or all five, in any combination, of the following techniques: stable isotope labeling of a substrate, spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction, to generate custom "NMR Metabolite Arrays" in which the resonances of all known metabolites of a given cell sample are categorized and used for comparison for simplified statistical analysis. This is the first time all these techniques have been combined to provide a robust, efficient and high throughput NMR metabolite screening protocol. The combinations of steps allow for global, unbiased, ultra-high resolution of both water-soluble and lipid-based metabolites. In addition, the novel "NMR Metabolite Array" programs described herein provide a new way to analyze large complex NMR datasets in a simplified manner. The new platform permits both the rapid identification of differentially expressed metabolites, quantification of specific metabolites, and the ability to analyze the metabolic flux of given precursors.

The new methods enable one to specifically follow the metabolic breakdown of a particular molecule in relatively few cells, e.g., about 2-20 million cells. The methods involve preloading cells with a labeled precursor substrate (e.g., labeled with ^{13}C, or ^{15}N, or ^{31}P) and using multidimensional NMR. The methods do not require purification of the individual metabolites of interest prior to analysis allowing for global, unbiased identification of metabolites that are differentially generated in cells with different properties.

Identification of metabolites differentially expressed in normal and disease state cells can be a powerful tool in the clinic. The new methods for monitoring differential expression of metabolites from cells that are phenotypically different are particularly useful for identifying therapeutic targets that can be used to modulate the phenotype. This includes targets that are present in the biosynthetic pathway of the metabolite, or the metabolite itself. Further, identifying differentially expressed metabolites can be used to differentiate cells of a normal versus a disease state. Hence, they have the potential to serve as a biomarker for the phenotype with which it is associated, making the methods described herein useful for identifying diagnostic markers, e.g., markers for diagnosis of disease.

For example, described herein we identified N-acetylneuraminic acid (NANA) as a novel biomarker for breast cancer tumor initiating cells, and monitoring its expression could be useful in diagnosing and detecting breast cancer. In addition, the protein level of CMAS, an enzyme in NANA biosynthesis was shown to be dramatically over-expressed in breast tumor initiating cells. Before this work, the role of CMAS in tumor initiation and metastasis has not been explored. Herein we provide evidence that CMAS expression is absolutely crucial for tumor formation and migration and that CMAS is a novel bona fide target for breast cancer.

The application of this methodology to an individual patient's cell analysis will also provide the basis for a "personalized medicine" approach to patient care.

Accordingly, this disclosure describes methods for monitoring the metabolism of a given substrate precursor within a cell population, e.g., a primary cell population, a tissue cell population, or cultured cells (e.g., immortalized cells). Using the methods described herein, the identification of differentially expressed metabolites between two or more cell populations that have different phenotypes is described. Also described are methods for identifying potential therapeutic targets and diagnostic markers.

In general, in a first aspect, the disclosure features new methods of monitoring metabolism of a substrate within a given type of cell in a sample. The new methods include (a) culturing a given type of cell of a first sample with a substrate for a sufficient period of time to allow metabolic breakdown of the substrate into substrate metabolites, wherein at least a portion of the substrate is optionally labeled with a nuclear magnetic resonance (NMR) stable isotope; (b) harvesting the substrate metabolites from the cells of step (a) to obtain a second sample of substrate metabolites; and (c) performing multi-dimensional NMR on the second sample of step (b) to determine a resonance spectrum of the metabolized substrate, wherein the resonance spectrum represents the metabolites of the substrate, and wherein the multi-dimensional NMR comprises any one of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

In another aspect, the disclosure features methods for identifying differentially expressed substrate metabolites between a first population of cells and a second population of cells. These methods include (a) optionally loading a first and a second population of cells

with a nuclear magnetic resonance (NMR) stable isotope-labeled substrate; (b) culturing the first and the second population of cells of step (a) for a sufficient period of time to allow metabolic breakdown of the substrate into substrate metabolites; (c) harvesting the substrate metabolites from the first and the second population cells of step (b) to obtain a sample of substrate metabolites from each of the first and the second cell populations; (d) performing multi-dimensional NMR on the sample of step (c) for each of the first and the second cell populations to determine a resonance spectrum of the metabolized substrate of the first population of cells and of the second population of cells, wherein the resonance spectrum represents the metabolites of the substrate; and (e) comparing the resonance spectrum of the first population of cells with the resonance spectrum of the second population of cells to determine which resonances are differentially expressed, wherein the differentially expressed resonances provide a resonance signature that represents differentially expressed metabolites.

In any of these methods, the multi-dimensional NMR can include any two, three, or all four, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction. In any of these methods, the substrate can be labeled with an NMR stable isotope and the multi-dimensional NMR can include any two, three, or all four, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

In some implementations of these methods, the substrate metabolites that are present in the sample are not purified away from the other molecules in the sample. In some implementations the substrate concentration within the population of cells is reduced for a period of time prior to loading the cells with the NMR-labeled substrate and the resonances of the metabolites of the labeled substrate are determined using NMR pulse programs or filtering techniques, or both, customized to the substrate. The number of cells within the population of cells can be is less than 2×10^6 and the population of cells can be a primary population of cells.

Any of these methods can further include comparing the resonance signature of step (e) with a database of known resonance signatures to determine the molecular structure that

the resonance signature represents, and thereby determine the substrate metabolites that are differentially expressed between the first and the second populations of cells.

In certain implementations, the methods can further include identifying a biosynthetic pathway involved in generation of the substrate metabolites and identifying proteins/enzymes of the pathway that may be targeted to modulate the differential expression of the metabolite, to thereby modulate the phenotype of the cells. In some embodiments the first population of cells and the second population of cells are isogenic populations and/or the first population of cells and the second population of cells have different phenotypes. In some implementations, the first population of cells is a control population of cells and the second population of cells has been contacted with a test compound or agent. The methods can be used to identify metabolic pathways that are overactive or underactive in a particular cell type. The methods can further include inhibiting or overexpressing a gene in the second population of cells and the method is used to identify the metabolic consequences of over-expressing or inhibiting a gene in a cell.

In another aspect, the disclosure features methods for treating cancer in a subject. The methods are based on results determined using the new platform methods described herein. The methods of treating cancer include administering to a subject in need thereof an effective amount of an inhibitor of N-acylneuraminate cytidylyltransferase (CMAS), an inhibitor of N-acetylneuraminic acid synthase (NANS), or a molecule that decreases the expression of N-acetylneuraminic acid. For example, the inhibitor can be an enzyme or can be selected from the group consisting of a small molecule, a ribonucleic acid, a deoxyribonucleic acid, a protein, a peptide, and an antibody.

As used herein, the term "isogenic" refers to cells of the same genetic background (any cell type, e.g., epithelial, or fat, or stem, or muscle cells etc.) that are isolated from the same tissue type (any tissue, e.g., tissue of the same organ, skin, bladder, liver, heart, etc.) and from the same organism type (e.g., human, or animal, or fish).

As used herein, the term "metabolite" refers to the intermediate or the end products of metabolism. "Metabolites" have functions comprising energy source, structural, signaling, stimulatory, and inhibitory effects on enzymes. Metabolites can also have catalytic activity themselves. A metabolite can be the end product of a substrate-enzyme reaction.

As used herein, the term "metabolic precursor" is a compound that participates in a chemical reaction. The term is meant to include to a compound that is a starting compound or an intermediate compound of an enzymatic reaction from which an end product results.

The term "substrate" refers to a molecule or compound on which an enzyme acts and results in the substrate transforming into one or more end products. The end products are released from the active site of the enzyme.

The term "enzyme" refers to a molecule that accepts a substrate in its active site and transforms the substrate into one or more end products that are subsequently released from the active site.

As used herein, the term "primary cell" or "primary tissue" refers to cells or tissue taken directly from living tissue of a normal individual or an individual with an acquired or inherited disease and established to grow *in vitro*.

The term "metastasis" refers to a process by which cancer spreads from the place at which it first arose as a primary tumor to distant locations in the body as well as the newly established tumor itself, which is also referred to as a "metastatic tumor" that can arise from a multitude of primary tumor types, including but not limited to those of prostate, colon, lung, breast, bone, and liver origin. Metastases develop, e.g., when tumor cells shed from a primary tumor adhere to vascular endothelium, penetrate into surrounding tissues, and grow to form independent tumors at sites separate from a primary tumor.

The term "cancer" refers to cells having the capacity for autonomous growth. Examples include cells having an abnormal state or condition characterized by rapidly proliferating cell growth. The term is meant to include cancerous growths, e.g., tumors (e.g., solid tumors); oncogenic processes, metastatic tissues, and malignantly transformed cells, tissues, or organs, irrespective of histopathologic type or stage of invasiveness. Also included are malignancies of the various organ systems, such as respiratory, cardiovascular, renal, reproductive, hematological, neurological, hepatic, gastrointestinal, and endocrine systems; as well as adenocarcinomas which include malignancies such as most colon cancers, renal-cell carcinoma, prostate cancer and/or testicular tumors, non-small cell carcinoma of the lung, cancer of the small intestine, and cancer of the esophagus. Cancer that is "naturally arising" includes any cancer that is not experimentally induced by implantation of cancer cells into a subject, and includes, for example, spontaneously arising

cancer, cancer caused by exposure of a patient to a carcinogen(s), cancer resulting from insertion of a transgenic oncogene or knockout of a tumor suppressor gene, and cancer caused by infections, e.g., viral infections. The term "carcinoma" is art recognized and refers to malignancies of epithelial or endocrine tissues. The term also includes carcinosarcomas, which include malignant tumors composed of carcinomatous and sarcomatous tissues. An "adenocarcinoma" refers to a carcinoma derived from glandular tissue or in which the tumor cells form recognizable glandular structures.

As used herein, the term "treating" or "treatment" refers to administering one or more of the compounds described herein to a subject who has an a disorder treatable with such compounds, and/or a symptom of such a disorder, and/or a predisposition toward such a disorder, with the purpose to confer a therapeutic effect, e.g., to cure, relieve, alter, affect, ameliorate, or reduce the disorder, the symptom of it, or the predisposition.

As used herein, the term "an effective amount" or "an amount effective" refers to the amount of an active compound that is required to confer a therapeutic effect on the treated patient. Effective doses will vary, as recognized by those skilled in the art, depending on the types of diseases treated, route of administration, excipient usage, and the possibility of co-usage with other therapeutic treatment.

Dosage, toxicity, and therapeutic efficacy of therapeutic compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD50 (the dose lethal to 50% of the population) and the ED50 (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD50/ED50. Compounds that exhibit high therapeutic indices are preferred. While compounds that exhibit toxic side effects may be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references

mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the inventions will be apparent from the following detailed description, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig.1 is a flow chart summarizing the key steps of the NMR based metabolite screening platform described herein.

Figs. 2A-F detail each step of the acquisition and processing in the NMR methods that allow for rapid, unbiased, ultra-high resolution metabolite profiling. Figs. 2A-2D demonstrate spectral width (sw) folding strategies, in which decreasing the sw from 220 pm to 90 ppm led to 2.5 fold increase in resolution. Fig 2E summarizes a non-uniform sampling (NUS) strategy allowing either an 8-fold increase in resolution or a 4-fold increase in resolution in time reduction by 40%. Fig. 2F shows a NUS 13C-1H HSQC spectra (left) processed using forward maximum entropy reconstruction and the same NUS 13C-1H HSQC spectra (right) after including a data-extension step in the reconstruction that increased resolution by 2-fold.

Figs. 3A and 3B are ^{13}C-^{1}H HSQC spectra indicating the full metabolic coverage of water-soluble (3A) and lipid-based (3B) metabolites from the same p53 deficient mouse lung tumor, respectively.

Fig. 4 is a flow chart summarizing the custom NMR analysis program used to create NMR arrays for rapid analysis.

Figs. 5A-C are a set of ^{13}C-^{1}H HSQC resonance spectra of water-soluble metabolites from 20 million unlabeled (5A) ^{13}C-gluatmine incubated (5B) and ^{13}C-glucose incubated (5C) breast tumor initiating cells.

Fig. 6 highlights the information available in the NMR arrays described herein. Using the unlabeled, glutamine, and glucose spectra in Figs. 5A-C, a master look up was created to generate metabolite IDs for all possible resonance metabolites that and these are listed on the X-axis. The Y-axis displays the relative intensity of each resonance in each condition, highlighting the differential glucose and glutamine derived metabolites.

Figs. 7A-B are representative examples of ^{13}C-^{1}H HSQC resonance spectra of water soluble glucose derived metabolites in breast tumor initiating BPLER cells (Fig. 7A) and less malignant isogenic HMLER cells (Fig. 7B).

Fig. 7C summarizes the NMR arrays for the BPLER and HMLER ^{13}C-^{1}H HSQC resonance spectra, showing how the intensity (Y-axis) of all possible metabolite resonances (X-axis) changes in each cell type. BPLER and HMLER cells originate from the same normal breast tissue and were grown into two cell types BPECs (breast primary epithelial cells) grown in chemically defined WIT medium and HMECs (human mammary epithelial cells) grown in MEGM media. BPEC and HMEC cells were transformed with hTERT (L), the SV40 early region (E), and H-ras (R) to give rise to BPLER and HMLER cells.

Fig. 7D is a zoomed in region of the a overlay of BPLER (red) and HMLER (blue) ^{13}C-^{1}H HSQC spectra in a region the NMR array predicted to only have BPLER resonances.

Figs. 8A-C summarize various methods used to validate that the differentially expressed metabolite overexpressed in the BPLER NMR arrays is N-acetylneuraminic acid (NANA). Fig 8A is the ^{13}C-^{1}H HSQC of pure NANA. Fig. 8B illustrates the results of custom NMR HCN (hydrogen-carbon-nitrogen) experiment, confirming BPLER cells have a differentially expressed resonance with similar connectivity to NANA. Fig 8C shows the M/S results directly measuring NANA in HMLER (left) and BPLER (right) cells using liquid chromatography-mass spectrometry multiple reaction monitoring (LC/MS MRM) where the number reported is the area under the NANA peak.

Fig. 9 is a series of representations of microscope images of rhodamine-labeled wheat germ agglutinin (WGA) immune-fluorescent microscopy showing the expression of NANA in HMLER (top row) and BPLER cells (bottom row), respectively. WGA specifically binds to NANA.

Fig. 10 is a schematic diagram depicting the enzymatic steps to convert glucose to NANA, in which N-acetylneuraminic acid synthase (NANS) and N-acylneuraminate cytidylyltransferase (CMAS) are key enzymes.

Fig. 11 is a bar graph showing the effects on proliferation by downregulating CMAS, NANS, and PLKI via siRNA on the viability of HMLER and BPLER cells.

Figs. 12A-C are a series of images that provide an overview of the NANA effect on cell migration of BPLER and HMLER cells. Figs. 12A and 12B are each a series of

immunohistochemistry images showing the inhibition of cell migration in the absence of NANS and CMAS and the rescue of migration with NANA in HMLER and BPLER cells, respectively. Fig. 12C is a bar graph quantifying the number of migrating cells in the absence of NANS and CMAS and the subsequent migration following addition of NANS.

Figs. 13A-B are a bar graph depicting the quantitative PCR (mRNA levels) of CMAS and cMYC (13A) and a Western Blot analysis of NANS and CMAS expression (13B), respectively, in HMLER and BPLER cells.

Figs. 14A and 14B are immunohistochemistry images of the effect of CMAS expression on cell migration. Fig. 14A is a series of immunohistochemistry images of HMLER cell migration following overexpression of CMAS. Fig. 14B is a series of immunohistochemistry images of BPLER cells and with stable CMAS knockdown BPLER cells (BPLER-shCMAS1).

Figs. 15A-B summarize the strategy to determine the effect of CMAS levels on tumor initiation and metastasis *in vivo*. Fig. 15A shows the immunization scheme, and Fig. 15B shows the tumor volume growth per day after in NOD/SCIN mice injected with 500,000 BPLER (left) or 500,000 BPLER-shCMAS1 cells.

Figs. 16A, 16B, 16C, and 16D are the enzyme mechanism of CMAS, a synthesized substrate based inhibitor of CMAS based on the structure of NANA, and immunohistochemistry analysis showing the inhibition of cell migration with the synthesized fluorine-NANA inhibitor, respectively.

Figs. 17A and 17C are the chemical structures of (2*R*,3*R*,4*S*)-4-guanidino-3-(prop-1-en-2-ylamino)-2-((1*R*,2*R*)-1,2,3-trihydroxypropyl)-3,4-dihydro-2*H*-pyran-6-carboxylic acid (which is Zanamivir, marketed as Relenza®)(Fig. 17A) and ethyl (3R,4R,5S)-5-amino-4-acetamido-3-(pentan-3-yloxy)-cyclohex-1-ene-1-carboxylate (which is oseltamivir, marketed as Tamiflu®)(Fig. 17C).

Figs. 17B and 17D illustrate the immunohistochemistry analyses of the effects of Relenza® on BPLER cell migration (17D) and a control (17B). Both of these drugs are neuraminidase inhibitors.

Figs. 18A and 18B are immune-fluorescent microscopy images showing the expression of NANA in HMLER and BPLER cells in the absence and presence of

neuraminidase (18A) and immunohistochemistry images of HMLER and BPLER cell migration in the presence and absence of neuraminidase (18B).

DETAILED DESCRIPTION

The present disclosure describes novel methods that provide benefits over numerous aspects of known NMR data acquisition and allow for rapid, unbiased, global, quantitative ultra-high resolution NMR data acquisition and custom NMR analysis utilizing a novel approach. The present disclosure describes new NMR screening methods that can be used to identify, follow, and characterize the metabolic breakdown of a particular molecule and to analyze the cellular metabolomics essential to a given cell type. The methods described herein circumvent the hurdles presented by known NMR protocols, namely a reduction in the sample size necessary to perform the data acquisition, eliminating the need for purification of the metabolites to be analyzed, reduced experimental time required for multi-dimensional NMR and obtaining high resolution necessary for metabolite identification. The method requires relatively few cells (2- 20 million), allowing the methods to be used to study metabolites from primary cells and tissues rather than just from cell cultures. In addition, the disclosed methods do not require purification of the individual metabolites of interest from other cellular metabolites prior to analysis. Further, the new methods allow one to visualize the specific metabolic fate of a given precursor in any cell type, and thus provide for the simplified identification of metabolites that are differentially generated in different types of cells utilizing novel data analysis methods also described herein.

To highlight the power and breadth of the new platform methods, this disclosure describes the identification of differentially expressed metabolites in triple negative breast cancer tumor-initiating BPLER cells that are highly aggressive compared to a less malignant isogenic line HMLER. It is widely known in the field that cancer cells thrive on glucose consumption. The glucose metabolite N-acetyl-neuraminic acid (NANA) has been identified herein as being differentially expressed in breast cancer tumor-initiating BPLER cells (i.e., increased expression). The biosynthetic pathway that generates the metabolite has also been identified and used to identify proteins/enzymes required for the synthesis of the metabolite as candidate targets within the pathway to modulate the phenotype of increased tumor initiation and metastatic potential.

In addition, the results of knock-down experiments in which genes that produce key enzymes NANS and CMAS are silenced demonstrated that the reduction of the normal function of these enzymes to generate NANA and attach it to proteins had no effect on the proliferation of BPLER cells, but greatly reduced their migration. On the other hand, forced over-expression of the same enzymes in HMLER cells increased their migration. Stable knockdown of CMAS in BPLER cells completely prevented tumor formation in mice. Thus, the new methods were successfully used to identify metabolites (e.g., NANA) important for tumorigencity and to demonstrate that NANA and the proteins involved in the synthesis of NANA may serve as targets for therapeutic intervention to reduce breast tumor formation and the metastasis of tumor-initiating cells. In addition, NANS, and CMAS were validated as targets for inhibition of tumor initiation and metastasis *in vitro* and *in vivo*. Thus, small molecule and other inhibitors of these enzymes are new candidate therapeutic agents that can be used to specifically target breast tumor-initiating cells. Furthermore, the differentially expressed metabolites (e.g., NANA) can also serve as biomarkers for the phenotype with which they are associated, allowing the methods described herein to be used to identify new candidate diagnostic markers.

General Methodology

In general, the new methods described herein for monitoring metabolites of a given precursor molecule (i.e., substrate) within a given type of cell include the steps of: (a) optionally loading a population of cells with a labeled substrate, e.g., a ^{13}C-labeled substrate, (b) culturing the cells of step (a) for a sufficient period of time to allow metabolic breakdown of the substrate into substrate metabolites (e.g., typically 5 minutes to 24 hours depending on the experimental question); (c) harvesting the substrate metabolites from the cells to obtain a sample of substrate metabolites, e.g., a water-soluble sample of substrate metabolites and organic sample of lipid-based metabolites; and (d) performing multi-dimensional NMR on the sample of step (c) to determine the resonance spectra of the metabolized substrate, wherein the resonances represents the resonances of metabolites of the substrate. How the multi-dimensional NMR is performed is described in further detail below.

In these methods, various substrates and various stable spin-1/2 nuclear isotopes can be used to label those substrates. For example, glucose, glutamine, fatty acids, amino acids,

pyruvate, drug compounds, and other molecules can be used as substrates, and stable isotopes such as ^{13}C, ^{15}N, ^{29}Si, ^{31}P, or others can be used to label the substrates. Any tissue, primary cells or cultured cell lines can be used for the analysis, such as cancer cells, muscle cells, fat cells, endothelial cells, epithelial cells, neuronal cells, cardiac cells, and many others. In general, one would want to test cells associated with a particular disease or disorder, such as a cancer cell, as well as the same type of cells from a healthy subject, to provide a differential metabolic analysis. One could also test cells with a single gene mutation (i.e. mutant cells vs. wild-type) or cells treated or not with a drug, or the same cells incubated with the precursor for various times. The cell populations of each sample can be a homogeneous cell population. In alternative embodiments, the cell population of each sample can be a heterogeneous cell population (e.g., derived from a tissue sample). While any number of cells can be used in the methods described herein, in various embodiments the number of cells within each population of cells can be less than 1 X 10^8, less than 8 x 10^7, less than 7 x 10^7, less than 6 x 10^7, less than 5 x 10^7, less than 2.5 x 10^7, or less than 2.0 X 10^7, or the cell number can range from approximately 1 x 10^6 to 1 x 10^8 cells, or 1 x 10^6 to 5 x 10^7, or 1 x 10^6 to 2.5 x 10^7, or 1 x 10^6 to 2.0 X 10^7.

Methods for loading cells are well known to those of skill in the art, e.g., the labeled substrate can be added to the cell culture medium for a period of time (e.g., 5 minutes to 24 hours), or may be loaded by transfection (e.g., liposome or calcium phosphate), transduction (e.g., viral delivery of labeled substrate) or transfusion (e.g. direct injection into a tumor). In one embodiment, labeled precursor is administered to a subject, e.g., orally, topically, parenterally, or intravenously. In some embodiments, the labeled substrate is added to animal feed.

In various embodiments, prior to adding the culture medium containing the labeled substrate, the cells are incubated with cell culture medium that lacks any form of the substrate, e.g., lacks the substrate in unlabeled or labeled form. The cells may be incubated in this medium for minutes to hours, essentially starving the cells of the substrate. This helps reduce background in the later analysis. The cells are then incubated with cell culture medium containing only labeled substrate, (e.g., 5 minutes to 24 hours). Any concentration of substrate can be used. In various embodiments, the concentration ranges from 1 ng/ml to >1 mg/ml. A skilled artisan can easily determine the best concentration to use by testing various

concentration ranges. After incubation the cells are washed briefly and immediately harvested to separate the metabolites. To harvest the metabolites, a simple chloroform extraction may be performed in order to obtain a water-soluble sample or a non-water soluble sample of metabolites present in the organic layer. No additional purification is required.

Several nuclear magnetic resonance (NMR) techniques may be used in the methods of the invention, preferably multidimensional NMR. For example, heteronuclear single quantum correlation (HSQC) spectroscopy, variations of HSQC, and other multidimensional NMR techniques can be used. Methods for performing multidimensional NMR (e.g., 2D NMR and/or ^{13}C-^{1}H HSQC NMR) are well known to those of skill in the art.

The resonance spectra of the metabolites of the labeled substrate can be determined using NMR pulse programs, which can be customized to the substrate. In general, NMR uses a static, homogeneous external magnetic field to polarize the NMR sample. This primary field is typically called the "B0" field, and it defines a reference axis for the NMR system. The NMR sample is magnetized in the direction of the B0 field by placing the sample in the B0 field for a period of time (e.g., minutes) and allowing the sample to reach a thermal equilibrium state. The primary B0 field also typically defines the resonance frequencies of the spin-1/2 nuclear species in the sample. For example, a stronger primary field generally increases the nuclear spins' resonance frequencies. The nuclear spins "precess" about the B0 field at their respective resonance frequencies. In most NMR systems, the nuclear spins have a resonance frequency in the radio frequency (rf) range.

In NMR experiments, the nuclear spins in the NMR sample are manipulated by applying a time-varying magnetic field at the nuclear spins' resonance frequency. In some instances (e.g., for low flip-angle pulses), high-intensity radio frequency (RF) pulses provide fast, precise control of the nuclear spins. High-intensity RF pulses have the benefit of shorter pulse times, which reduces the amount of decoherence that occurs during the pulse. In some instances (e.g., for high-flip angle pulses), high-intensity RF pulses provide less precision, for example, due to non-uniform power output over a frequency range of interest, due to spatial inhomogeneity in the RF field, or due to other considerations.

In some implementations, adiabatic pulses can provide more precise control of the nuclear spins. Adiabatic pulses typically have a lower intensity and require a longer pulse time. In some cases, adiabatic pulses are used for larger flip-angle pulses (e.g., 180 degree

flip angle) to provide a more uniform flip angle over the entire frequency range of interest. Adiabatic pulses are typically implemented shaped pulses (meaning that they have a time-varying power profile) that can be parameterized for a particular flip angle, a particular frequency range, etc.

In various embodiments, the new NMR methods include the use of any one or more the following methods and techniques to increase analysis speed and/or resolution, or both: (i) stable isotope labeling (ii) folding the spectra width and aliasing peaks, (iii) random phase sampling (iv) non-uniform sampling, and (v) data extension. In some embodiments the methods include at least two or three of these methods in any combination. In some embodiments, the methods include all five methods. Table 1 below summarizes the effects of each step on the NMR acquisition and data resolution. These steps are described in more detail below.

TABLE 1

METHOD	EFFECT ON NMR ACQUISITION & DATA RESOLUTION
Stable isotople labeling	Increases metabolite signal detection by 99%
Folding spectral width (sw)	Increases resolution by 2.5 fold to 44 Hz (approaching theoretical limit of C-C decoupling)
Random phase sampling	Decreases the total acquisition time by 50%
Non-uniform sampling	For time equivalent spectra one can gain 8-fold increase in resolution or one can gain 4-fold increase in resolution and at the same time cut experimental time by 40%.
Data-extension	2-fold increase in data resolution with no effect on experimental time

Stable Isotope Labeling

Traditional 2-D NMR metabolite profiling relies on ^{13}C-natural abundance, which exists at 1.1%. As such, in order to observe any signal at all, large amounts of sample are required (on the average +200 million cells). This limits detection to cultured cells lines and only the most abundant metabolites are detected. To decrease the amount of material needed and view metabolites with a broader concentration range, samples (e.g., cells, tissues or tumors) were directly supplemented with ^{13}C-labeled precursors (glucose, glutamine, pyruvate and amino acids were used but other substrates and other isotopes are also possible). Theoretically, this should decrease the sample burden by ~99%. If, for example a 1 mM metabolite required 200 million cells to be detected with ^{13}C natural abundance, using a label to see the same intensity would require only 2 million cells. Including this step in our method reduces sample requirements to those similar for some mass spectrometry (M/S) approaches. Untargeted M/S requires at least 1million cells and is followed by several rounds of chromatography purification and detection. Our method requires few cell numbers and no purification.

Folding the Spectra Width and Aliasing Peaks

When an atom is placed in a strong magnetic field (B0), the electrons in that molecule precess in the direction of the applied magnetic field. This precession creates a small magnetic field at the atomic nucleus. The magnetic field at the nucleus (B) is therefore generally less than the external magnetic field (B0) by τ.

$$B = B0(1-\tau).$$

The electron density around each nucleus within a molecule varies according to the types of nuclei and the bonds in the molecule. The opposing field and therefore the effective magnetic field at each nucleus will vary. In pulsed NMR spectroscopy these differences can be measured by applying a radio frequency pulse that causes the nuclear magnetization to oscillate inducing an electrical current in a coil that can be measured. This signal, known as "free induction decay" (FID) is plotted as current with respect to time. By applying a discrete

Fourier transform, the FID can be converted to frequency domain and the resonance frequency of each observable nuclei can be converted to chemical shift (δ) by the equation,

$$\delta = (n\text{-}n_{REF})x10^{\wedge 6}/n_{REF}$$

where n is the resonance frequency of the nucleus and n_{REF} is the resonance frequency of a standard.

Chemical shift is a very precise metric of the chemical environment around a nucleus. Unlike M/S and chromatography, NMR is one of the only methods that can distinguish molecules that have the same mass but different chemical connectivity. However to utilize this information ultra-high resolution spectroscopy is needed.

In NMR, digital resolution is determined by the sweep width (sw) and the total number of data points (TD), such that

Resolution = sw/TD, measured in Hz/point.

SW is the range of frequencies over which NMR signals are to be detected. Metabolite mixtures contain diverse molecules, and the spectral width necessary to cover all potential carbon chemical shifts spans over -220 ppm. The large ^{13}C-chemical shift window creates a dilemma, where in order to have maximum resolution and a broad enough sw to encompass all possible chemical shifts one would require an incredibly large number of data points. In practical terms this is time-prohibitive.

To circumvent this, our method can include folding the sw. The sw is purposefully set to smaller range and if some peaks occur outside this range they will appear "folded" at aliased chemical shifts. Folded spectra can be unfolded by suitable data processing techniques. For example, resonance frequencies can be dealiased by expanding the frequency spectrum and shifting the aliased frequencies by a pre-defined amount, to their actual locations. In some cases, the data acquisition parameters define the spectral folding windows in a manner that reduces or minimizes any overlap between folded spectral peaks and non-folded spectral peaks. As such, the folded spectral peaks can be de-aliased without affecting other data in the frequency spectrum, in some cases. Folding spectra decreases the overall

number of points required in order to achieve the maximum resolution possible. Our custom folding strategy and de-aliasing program allows ultra-high resolution spectra with ~44Hz/point separation.

Random Phase Sampling

As described above converting the FID signal to frequency data requires Fourier transformation. However, for a nuclei rotating at +x magnetization vector around the Z-axis, the Fourier transform will give peaks at both $+\nu$ and $-\nu$ because the Fourier transformation cannot distinguish between a $+\nu$ and $-\nu$ rotation of the vector. The most common method to distinguish the sign of the frequency requires sampling the signal at two different receiver phases (for example 0° and 90°). For multidimensional NMR this increases the experiment time by a factor of two for each dimension. To increase the speed of our analysis we employed random phase sampling (RPS) where a single phase is used to detect each point but the phase is randomly alternated for different points in the signal. This allows us to resolve the phase of the frequency but cut the acquisition time in half.

Non-Uniform Sampling and Data Extension

Two-dimensional NMR techniques generate two dimensions of data in the time domain: a direct domain and an indirect domain. The direct domain data are generated by running an experiment and collecting an NMR signal (e.g., an FID, an echo, or a stroboscopic signal). In other words, the direct domain is the time domain of an NMR experiment. The indirect domain data are generated by systematically varying a time parameter of the NMR experiment (e.g., incrementing a delay time), running the NMR experiment for each value of the parameter, and combining the NMR signals from all experiments. In other words, the indirect domain is the time domain of the parameter that is systematically varied.

In some cases, non-uniform sampling can be used in multi-dimensional NMR for the methods described herein. For example, non-uniform sampling can be used in the indirect domain to reduce the number of NMR experiments that are needed to obtain a particular spectral range and frequency resolution.

Non-uniform sampling (NUS) can be accomplished by incrementing the indirect domain time parameter systematically and in a non-uniform manner. In particular, instead of incrementing the time parameter by the same amount for each successive NMR experiment, the time parameter can be incremented by an amount that varies depending on one or more factors. For example, the time delay parameter can be incremented by an amount that changes (e.g., increases or decreases) from experiment to experiment. Varying the time delay according to a Poissonian distribution or another nonlinear distribution results in sparsely sampled indirect domain data. The missing points in the "sparse" data set can be calculated using reconstruction methods. The forward maximum entropy reconstruction technique can conserve the measured time-domain data points and guess the missing data points by an iterative process. The iterative process can include discrete Fourier transformation of the sparse time-domain data set, computation of the spectral entropy, determination of a multidimensional entropy gradient, and calculation of new values for the missing time-domain data points with a conjugate gradient approach. Since this procedure does not alter measured data points, it can reproduce signal intensities with high fidelity and avoid dynamic range problems. In some cases, our method indicates with appropriate sampling schedules NUS has enhanced ability to detect weak peaks. This is extremely important for metabolite analysis where there is a large dynamic range between abundant metabolites (milimolar concentration) and rare metabolites (nanomolar concentration).

During the reconstruction it is possible to further increase the resolution of resonances utilizing "data extension." In this method, during the reconstruction, the total number of points in the indirect dimension is doubled. In one embodiment, the first half of the time domain (composed of NUS sampled points) is solved according to standard forward maximum entropy protocols and the second half of the data that does not contain any sampled points is completely built using iterative soft thresholding. In some cases, this allows a two-fold increase in resolution without affecting acquisition time.

Optional Enhancements to Reduce Background

In general, when using the methods described herein there is no need to purify or isolate the substrate metabolites that are present in the sample from the other molecules in the sample (e.g., by chromatographic column (e.g., Sidelmann, et al. Purification and 1H NMR

Spectroscopic Characterization of Phase II Metabolites of Tolfenamic Acid Drug Metab Dispos June 1, 1997 25:725-731) or by other means well known to those of skill in the art). Thus, in typical embodiments, the only purification performed in the method prior to NMR is the separation of the metabolites into a water soluble sample or a non-water soluble sample.

In some embodiments, the concentration of the unlabeled substrate within the cell can be reduced for a period of time (e.g., 10 minutes to 4 hours) before adding the labeled substrate. This technique can help reduce background signal. The appropriate time frame can be determined by testing a range of conditions and monitoring background as compared to control cells not loaded with labeled substrate.

Automated NMR Analysis

NMR metabolite analysis is tedious and complex. To circumvent these problems, a custom "NMR Metabolite Arrays" program was created to automate the process. As shown in Fig. 4, spectra are automatically first phased, aligned, and normalized with a spike-in control. For example, a known material, such as 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), tetramethylsilane (TMS), trimethylsilyl propionate (TSP), 4,4-dimethyl-4-silapentane-1-ammonium trifluoracetate (DSA) or other NMR standard reference compound can be added into each sample, and its concentration used as a reference to enable relative quantification comparisons between samples. Next, using a custom automatic peak picking program, peak lists can be generated for each sample, where each resonance peak is converted into an X, Y coordinate with an intensity value. A "MASTER PEAK LIST" program that generates a "master" look up table for all the resonances in the spectra under investigation can then be run. This program reads all X, Y points from the individual peak list files, removes duplicates within defined tolerances, and writes the resulting set of peaks to a standard output such that all possible metabolite resonances under investigation are determined. Depending on the analysis it is possible to input the entire HSQC data from the Human Metabolite Database into the master peak-list. However, taking this approach requires longer computation times, and in most cases is unnecessary. Thus, creating master look up tables for the spectra that are specifically being investigated is preferred.

As further shown in Fig. 4, after creating the master look up table, "NMR arrays" are next generated for each sample. NMR arrays consist of a list of all possible metabolites and intensity values for each resonance under investigation. They are created by combining the

individual test peak list and master peak list to fill in the intensity for resonances for all possible metabolites. If a metabolite is expressed in a test sample, the program will select that intensity value. If it is not present, the intensity is set at zero or an arbitrary number. The NMR arrays can then be analyzed via traditional statistical analysis programs to identify the differentially expressed resonances between spectra. The resonance frequencies can then be uploaded directly into a database, such as the Human Metabolome Database, to identify which metabolites are differentially expressed. Candidate metabolites can then be confirmed via additional NMR or M/S experiments.

Differential Expression Analysis

Also provided herein are novel methods for identifying substrate metabolites differentially expressed by at least two populations of cells, e.g., a first population of cells and a second population of cells. One of the populations of cells can be from a healthy subject or cell line and used as a control. The methods can use the various features of the various method steps and techniques described herein and include: (a) loading a first and a second population of cells with a labeled substrate (e.g., a ^{13}C, ^{15}N, or ^{31}P-labeled substrate); (b) culturing the first and the second population of cells of step a) for a sufficient period of time to allow metabolic breakdown of the labeled substrate into substrate metabolites; (c) harvesting the substrate metabolites from the first and the second population cells of step (b) to obtain a sample of substrate metabolites from each of the first and the second cell populations, (d) performing multidimensional NMR on the samples from each of the first and the second cell populations to determine the resonance spectra of the metabolized substrate, wherein the resonance spectra represents the metabolites of the substrate; and (e) processing the resonance spectra using a custom "NMR arrays" program (f) comparing the resonance intensity of the first population of cells with the resonance spectra of the second population of cells to determine which resonance spectra are differentially expressed, wherein the differentially expressed resonance spectra represents differentially expressed metabolites.

In some embodiments, the methods for identifying differentially expressed substrate metabolites between at least two populations of cells can further include comparing the resonance spectra of step (f) with a data base of known resonance spectra to determine the molecular structure(s) that the resonance spectra represents, and thereby determine which

specific substrate metabolites are differentially expressed between the first and second, different population of cells. Specific metabolites may be identified in this manner.

The methods described herein are useful for monitoring metabolism in any cell type, as well as in any tissue (e.g., the cell population can contain a heterogeneous population of cells), and are useful for monitoring metabolism in cells exhibiting any phenotype as compared to a cell not exhibiting the phenotype.

Diagnostic Applications

Identified differentially expressed metabolites are indicative of the different phenotype, and their expression can therefore be used to diagnose this phenotype, e.g. to diagnose increased metastatic potential, or to diagnose insulin resistance, or to diagnose disease, etc.

We discovered N-acetylneuraminic acid (NANA) is more highly expressed in breast cancer tumor initiating cells. From a diagnostic point of view, the presence of excess NANA can serve as a biomarker for tumorigenic potential. Once such differential expression is observed, detection methods such as antibodies or mass spectrometry can be utilized to monitor NANA expression (e.g., intracellularly or extracellularly) to help identify aggressive tumors. In summary, using the methods described herein for breast cancer cells, the specific breakdown of glucose was followed and NANA was discovered to be widely upregulated in more malignant cells. Enzymes were identified in NANA biosynthesis as new therapeutic targets, and the expression levels of the molecule was found to correlate with increased migratory potential.

The novel methods can also be applied to patient biofluids (e.g., blood, urine, plasma, and tissue samples) to discover metabolic differences that can serve as novel biomarkers for a particular disease.

Methods of Determining Candidate Therapeutic Agents

In various embodiments, the methods for identifying substrate metabolites differentially expressed by at least two populations of cells can further include the step of identifying a biosynthetic pathway involved in the generation of the substrate metabolites and identifying proteins/enzymes of the pathway that can be targeted to modulate the differential

expression of the metabolite. In turn, these proteins and enzymes can serve as candidate targets for modulating the phenotype of the cells, e.g., a disease phenotype, metastatic potential, or resistance. Thus, the new methods provide a means for identifying therapeutic targets. Once the metabolite is identified and an NMR metabolite array has been created for a given sample, databases of biosynthetic pathways can be screened to identify the pathway of synthesis of the metabolite.

For example, the NMR metabolite array can be electronically linked to the Human Metabolome Database and/or ChemPub, to select possible therapeutic targets within the metabolite biosynthetic pathways and propose substrate-based inhibitors using the metabolite itself as a lead scaffold for drug design. A series of differentially expressed metabolites that can serve as biomarkers have been identified and described herein. Novel therapeutic targets within the biosynthetic pathways as well as FDA-approved drugs that show efficacy in the laboratory that could be rapidly translated into new applications in the clinic have also been identified.

The cell populations used in the various methods described herein can be from healthy or diseased subjects. The cells can be from isogenic populations. In some embodiments, the first population of cells and the second population of cells have different phenotypes (e.g., differ in metastatic potential, differ in response to insulin, or differ in expression of disease genes). Differential metabolites expressed in any phenotype can be assessed as compared to an isogenic cell that does not exhibit the phenotype. Phenotypes are easily identified by those skilled in the art, and include but are not limited to phenotypes associated with a particular disease or disorder.

In some embodiments, the first population of cells is a control population of cells and the second population of cells has been contacted with a test compound or agent (e.g., after treatment with the compound or agent). The disappearance of differentially expressed metabolites that are associated with a particular phenotype serves as an indicator that the test compound or agent is capable of inhibiting the phenotype, for example, inhibiting metastasis, or inhibiting the effects of the expression of a diseased gene. Alternatively, the appearance of a differentially expressed metabolite (e.g., one expressed only in normal cells as opposed to diseased cells) serves as an indicator that the compound or agent is useful for treatment of the disease. Thus, in various embodiments, the methods for identifying differentially

expressed metabolites can be used to screen for compounds or agents that modulate a phenotype, which can be used for treatment of disease.

The metabolic consequences of overexpressing or inhibiting a gene of interest can be identified using the new methods described herein. In one embodiment, the method further comprises inhibiting or overexpressing a gene in one of the cell populations (e.g., the second population of cells). Similarly, the metabolic consequences of particular compounds or agents, e.g., to assess toxicity, also can be identified.

The test compounds or agents can be, for example, a small molecule, a nucleic acid RNA (e.g., siRNA or microRNA), a nucleic acid DNA, a protein, a peptide, or an antibody. The inhibitors can be selected from the group consisting of: a small molecule, a nucleic acid RNA (e.g., siRNA), a nucleic acid DNA, a protein, a peptide, and an antibody. In one embodiment, the inhibitor is an inhibitor of an enzyme (e.g., a neuraminidase inhibitor).

Methods of Treating Disorders with Therapeutic Agents

As described in the examples below, through the methods described herein, CMAS, NANS (also known as sialic acid synthase), and NANA cell-surface expression have been determined to be therapeutic targets that decrease migration of cancer cells and prevent tumor initiation *in vivo*. Thus, in another aspect, the present disclosure includes new methods for treating disorders such as cancer (e.g., by inhibiting metastasis and/or blocking tumor initiation) in a subject by administering an effective amount of an inhibitor of the targets discovered using the new methods described herein. In particular, the methods include administering to a subject in need thereof a therapeutically effective amount of an inhibitor of CMAS or NANS, or a therapeutically effective amount of an inhibitor or agent that lowers NANA expression. For example we have identified candidate CMAS inhibitors including a molecule we designed and synthesized, termed F-NANA, and already FDA approved influenza drugs, including Relenza and Tamiflu. The efficacy of these drugs in *in vivo* mouse models is being tested and because Relenza and Tamiflu have already been evaluated for safety in human subjects accelerated FDA approval for an investigational new drug is possible.

Application of the New Methods to Personalized Medicine

As mentioned, the new methods described herein have been successfully shown to characterize the metabolic differences in several oncology models using both cell lines and primary tissue. The methods have the potential to profoundly affect the strategies for designing novel therapeutic intervention and could be lay the foundation for a metabolite-based approach for "personalized medicine."

According to the National Institutes of Health, "personalized medicine" is a practice of medicine that uses an individual's genetic profile to guide customized decisions made with regard to the prevention, diagnosis and treatment of disease in that individual. To date, most efforts rely on genomic information to identify DNA mutations, amplifications, or deletions. Rarely, however, is a disease the result of a single genetic lesion, and it is often not obvious how genetic variations will manifest themselves. However, non-genetic changes, including epigenetic differences, can also have profound effects on gene expression and cellular properties. Further, many commonly mutated genes, such as in cancer, do not have small molecule inhibitors and are often termed "undruggable targets." Establishing the metabolic profile of key cells in an individual suffering from a disease, with the methodology described herein, could provide powerful information useful in diagnosing, treating and monitoring the disease state of an individual.

As noted, diseases are incredibly complex and heterogeneous, and the effect of a misregulated gene or genes is not always obvious; making it difficult to design the best therapeutic strategies for intervention. Metabolism on the other hand is the end product of the genome. Using the new platforms described herein, the metabolic differences in a specific patient can quickly highlight functioning or non-functioning pathways in that individual. Further, metabolic pathways have been extensively studied and in many cases inhibitors for metabolic enzymes, such antimetabolites, substances bearing a close structural resemblance to the natural metabolite, already exist and are already used in the clinic. Examples of such potential therapeutic discoveries based on the differential expression NMR analysis are also presented in detail below in the Examples. The aforementioned differential expression, for example of metabolites, can be a powerful tool for diagnosing, monitoring, and treating disease in a patient on an individual, customized basis. Examples of analysis

using this protocol to deduce novel metabolism pathways and differential metabolite expression between wildtype and disease tissue are described herein.

EXAMPLES

The invention is further described in the following examples, which do not limit the scope of the invention described in the following examples, which do not limit the scope of the invention described in the claims.

The following examples discuss the novel protocols and the subsequent novel analysis methods that are used to efficiently determine the differential expression of metabolites in normal and disease state tissue. The results of such differential expression are described herein and are shown to be utilized in the design and identification of potential small molecule therapeutics.

Example 1: Methodology for the Screening Platform Utilizing NMR

Preparation of Biological Sample: Figure 1 provides an overall description of the platform. Approximately 20 million cells were used for each sample (however it is possible to use as few as 2 million). Before harvesting, ^{13}C-labeled precursors (glucose and glutamine in these examples) were added directly to the media and allowed to incubate for a user-defined amount of a time and in this case 4hrs. After aspirating the media and washing two times with phosphate-buffered saline (PBS), the cells were again counted and collected. An equal number of cells with no label were also harvested to serve as a ^{13}C background control. Cells were lysed by the addition of ice-cold methanol, and an aqueous extraction was performed by adding equal parts water and chloroform. After centrifugation the water soluble and organic metabolites were separately collected, dried, and stored until ready to be further analyzed. No additional purification was performed.

Acquisition of Data with NMR: 2D NMR spectroscopy was employed primarily relying on heteronuclear single quantum correlation (HSQC) to identify metabolites. HSQC experiments provide one-bond correlation between a heteronucleus (^{13}C in the following examples, although other isotopes are feasible) and a proton. Crosspeaks arise due to transfer through the relatively large one-bond heteronuclear coupling, making it possible to identify shifts of directly attached nuclei. The unique chemical environment of each carbon atom

paired to a proton gives rise to characteristic chemical shifts specific to a given metabolite. Reference HSQC spectra of purified metabolites (commonly available) were used for comparison. For example, at present the Human Metabolome Database (HMDB) contains information on 40,260 metabolite entries many with HSQC data.

To overcome the traditional drawbacks of 2D NMR metabolite profiling (large sample requirements and long acquisition times), several additional techniques were used to improve the resolution and reduce the time required for the analysis. As described above, in a first step, to decrease the amount of material needed cells were supplemented with ^{13}C-labeled precursors (glucose, glutamine, pyruvate and amino acids were used but other substrates and other isotopes are also possible). Theoretically, this should decrease the cell number required by a factor of ~100. Partly due to varying ionic strengths, this does not always scale perfectly linearly and if there are no constraints regarding the sample size (i.e., using cell lines), it is recommended here that about 2-20 million cells be used for each analysis.

Example 2: Rapid Ultra High Resolution NMR Data Acquisition

Folding the spectra: While the aforementioned sample preparation alleviated the physical demands on the amount sample, the long acquisition time required to record high resolution 2D NMR spectroscopy was still a concern. To combat this, a multi-prong approach was taken: "folding" the spectra width, using random phase sampling (RPS), implementing non-uniform sampling (NUS) techniques and data extension in the analysis.

As described above, the spectral width (sw) is the range of frequencies over which NMR signals are to be detected. Metabolite mixtures contain diverse molecules, and the spectral width necessary to cover all potential carbon chemical shifts spans over -220 ppm. In Figure 2A, the HSQC spectra for all water soluble metabolites in KRAS mutant pancreatic cancer cells is shown. In this instance the 13C-sw spans 220 ppm and a total 1024 points were collected. By solving the equation, resolution = SW/TD (where TD is the total data points), we observed the resolution is limited to ~107Hz/point. However, if you examine Figure 2A closer, the majority of metabolites run along a diagonal and most the spectrum is empty. Collecting points along the entire sw greatly diminishes resolution, and as acquisition time is the reciprocal of resolution, experimental time is also wasted. As such we gradually

decreased the sw to increase the resolution. Figure 2B shows an HSQC of the same sample with 140 ppm sw and as a result resolution of ~68 Hz/point, Figure 2C with sw of 110 ppm and as a result 54Hz/point resolution and Figure 2D with 90 ppm sw generates ultra-high resolution of ~44Hz/point. In each of the folded spectra the aliased peaks are easily identifiable and the true chemical shifts can be can be back calculated using the following equation ($\delta obs = \delta + sw$). Of note, the maximum resolution due to C-C scalar coupling is ~35Hz. Using our folding strategy we are able to obtain ultra-high resolution spectra.

Of note, to provide optimal flip angles uniformly across the large carbon spectral width, broad band adiabatic shaped pulses were utilized for all 180 degree pulses along the carbon channel. This is especially important for enabling efficient coherence transfer among scalar coupled spins.

Non uniform sampling and extension of data: The measured "free-induction decay" (FID) of an NMR sample is created by the oscillating current generated by the precession of all magnetized bonds. This signal decays due to nuclei in other molecules creating spin-spin decoherence. The rate at which this occurs is known as the transverse relaxation rate (T2). For any NMR experiment it is widely viewed that to obtain maximum resolution one should collect points in the indirect dimension close to 1.2*T2. However, metabolites move rapidly with molecular motion correlation times on the average of 10^{-12} to 10^{-11} sec. Due to this rapid movement, for many metabolites there is little spin-spin decoherence and the T2 rates are almost infinitely long. Thus collecting ultra-high resolution metabolite data is theoretically possible but in practice it would require extremely long measurement times and in most experiments only a subset of data is collecting sacrificing resolution for speed.

By employing non-uniform sampling (NUS) techniques that are outlined in Figure 2E we were able to not only greatly increase the resolution of our data but also increase the speed at which we recorded high resolution spectra. For example using the same sample we could perform a uniformly sampled experiment with 128 indirect points in 5 hours with X resolution, using NUS we could collect 10% of 1024 points and in equivalent time to generate a spectra with 8X resolution, alternatively we could sample 10% of 512 points increasing the resolution by 4X and decreasing the acquisition time by ~40%.

The Poisson-gap distribution was selected for the sampling schedule followed by forward maximum (FM) entropy reconstruction. Metabolite mixtures contain molecules at

various concentrations, and this has been shown to be the most effective method in detecting weak peaks. In addition, to further enhance our resolution we created a "data-extension" add-on, in which before reconstruction the total number of points in the indirect dimension is artificially doubled. The first half of the NUS data set is reconstructed using the sparsely sampled data and filling in the missing points according to FM reconstruction. The second half of the data is completely built using iterative soft thresholding. As shown in Figure 2F this increased our resolution by 2-fold without affecting acquisition time.

Analysis of metabolites in the water and the organic layer: While it is not necessary to follow each step, this method allows for a full metabolic profile of both water soluble or organic metabolites. Figure 3 shows the HSQC of water based metabolites (Figure 3A) and organic based metabolites (Figure 3B) from the same ~2 million p53 deficient lung cancer cells and each experiment required only 1 hour of acquisition time. Equivalent spectra using 13C-natural abundance and standard NMR techniques for the same sample amount would require several days. Both spectra are extremely well resolved, making it easy to identify metabolite resonance peaks. Importantly many M/S methods have struggled to accurately detect lipids, using our method, as shown in Figure 3B, the resonances from organic molecules are readily identifiable.

Example 3: Analyzing metabolite NMR data

NMR Analysis: As summarized in Figure 4, a custom "NMR Metabolite Array" program was created to automate the NMR analysis process using the method herein described. First, the spectra are phased, aligned, and scaled with an internal control. 1 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) was added into each sample, and its concentration was used as a reference to allow relative quantification comparisons between samples to be made. We created an automatic peak picking program to generate peak lists for each sample, where each resonance peak was converted into an X, Y coordinate with an intensity value. Next we created a "MASTER PEAK LIST" program that generates a "master" look up table for all the resonances in the spectra under investigation and was subsequently run.

In short, this program reads all X, Y points from the individual peak list files, removes duplicates within defined tolerances and writes the resulting set of peaks to a

standard output. Depending on the analysis it is possible to input the entire HSQC data from the Human Metabolite Database into the master peak-list. Taking this approach requires longer computation times, and in most cases is unnecessary. Creating master look up tables for the spectra that are specifically being investigated is preferred.

Next, NMR arrays are generated for each sample in which the individual peak list and master peak list were combined to fill in the intensity for resonances for all possible metabolites. If a metabolite is expressed in a test sample, the program will select that intensity value. If it is not present, the intensity is set at zero or an arbitrary number. The NMR arrays can now be analyzed via traditional statistical analysis programs to identify the differentially expressed resonances between spectra. The resonance frequencies can then be uploaded directly into the Human Metabolome Database to identify which metabolites are differentially expressed. Candidate metabolites can then be confirmed via additional NMR or M/S experiments.

Example 4: Analysis of Differentially Expressed Metabolites: A Novel Approach

Background Correction: To monitor the flux of a given precursor, a separate spectra with an equal number of cells with no labeled precursor was recorded. The spectra from the unlabeled cells represent the ^{13}C background within the cell and can be subtracted from the test spectra to specifically follow the metabolic breakdown of the ^{13}C labeled substrate. Glucose and glutamine are two of the main energy sources within a cell, and the metabolic breakdown of each precursor is well characterized. To examine the flux of glucose and glutamine into specific pathways, ^{13}C-^{1}H HSQC spectra of equal number of breast tumor initiating cells with no labeled precursor and either ^{13}C-glutamine or ^{13}C-glucose added as substrate were recorded. As shown in Figure 5A with this limited number of cells, very few signals arise from the ^{13}C background, and the glutamine (5B) and glucose (5C) spectra are quite distinct. By creating NMR arrays we were able to convert the complex NMR data into standard text files. The intensity values for resonances in the unlabeled sample were subtracting from the matching signals in the glutamine or glucose arrays. As shown in Figure 6, by plotting the intensity value and resonance metabolite ID from the NMR array it is possible to identify changes specific to glucose or glutamine flux through a given sample. The X-axis lists all the resonance metabolite ids for every metabolite identified from the

HSQC spectra of Figure 6b and Figure 6c. The Y-axis highlights how the intensity of each resonance changes in each condition. As expected, it is clear glutamine and glucose cause flux into different metabolic pathways. The resonance metabolite IDs correspond to specific ^{13}C-^{1}H chemical shifts which can uploaded into the Human Metabolome Database to identify the differential metabolites.

Example 5: Identifying differentially expressed metabolites in triple negative breast cancer tumor initiating cells

Protocol: Originating from the same normal breast tissue, BPLER and HMLER cells were transformed with identical genetic factors but were propagated in different culture media. BPLER are highly tumorigenic and have an increased metastatic potential over that of HMLER cells. Less than 50 BPLER cells injected into the mammary fat pad of a mouse result in the development of a tumor, while more than $10^{\wedge 6}$ HMLER cells are required to form a tumor *in vivo* (Table 2, below). BPLER cells are a model cell line for triple negative breast cancer tumor initiating cells, and BPLER tumors histologically resemble that of triple negative breast cancer patients. According to the protocol about 20 million BPLER and HMLER cells were cultured in the presence of uniformly labeled ^{13}C-glucose, and subsequently harvested and lysed. The aqueous layer was then collected, dried, and re-dissolved in ultra-pure D2O and ready for NMR analysis. The organic layer was stored for future examination.

TABLE 2

	Tumors Formed		
Cells	**BPLER**	**HMLER**	**MCF7**
$5x10^4$	4/4	0/4	0/4
$5x10^3$	4/4	0/4	0/4
$5x10^2$	4/4	0/4	0/4
5x10	4/4	0/4	0/4

Using the new platform methodology described herein, the rapid, unbiased, ultra-high resolution NMR metabolite screening was performed. Examples of resulting ^{13}C-^{1}H HSQC for BPLER and HMLER cells are shown in Figure 7A and 7B respectively.

Results: Using our custom NMR analysis program, the resonances in each spectra were converted into NMR arrays. Figure 7C summarizes the information. By combining all replicates from both cell lines approximately ~2100 resonances were identified. The metabolite ID of each resonance is listed on the X-axis. The relative intensity of each resonance is plotted on the y-axis. From this analysis we observed a high degree of similarity between the metabolite resonances of each cell line (>75% of the resonances were present in both cell lines). However several resonance peaks that were common to both cell lines had varied expression, while some peaks were unique to HMLER cells and others specific to BPLER. To confirm the arrays accurately reflected the HSQC data, as shown in Figure 7D the region of the HSQC spectra that was predicted to have resonances specific to BPLER cells in the NMR array was expanded and indeed resonances were only found in the BPLER spectra.

Using the NMR arrays we were able to quickly identify resonances that were specifically enriched in BPLER tumor initiating cells. Table 3 highlights the top resonances most enriched in BPLER tumor initiating cells. Shown are the metabolite IDs from the array, as well as the corresponding ^{13}C-^{1}H data.

TABLE 3

Metabolite Resonance ID	13C	1H
589	62.993	3.842
1951	65.98	3.823
283	68.633	4.248
1381	69.973	4.002
402	42.028	2.193
1602	72.951	3.739
245	72.849	3.953
1333	42.015	1.814
119	106.501	6.151

These resonances were input into the Human Metabolome Database and 6 of the 9 resonances, highlighted in yellow were predicted to be from N-acetylneuraminic acid

(NANA), strongly suggesting NANA is the metabolite corresponding to the differentially expressed resonances identified in the NMR arrays.

Several additional steps were taken to confirm NANA is indeed upregulated in BPLER tumor initiating cells. First, ^{13}C-^{1}H HSQC of pure NANA shown in Figure 8A contains cross peaks at approximately the same location as those found over-represented in the BPLER spectra. Second, we designed custom NMR pulse programs to specifically examine NANA. In NANA biosynthesis the C2 of glucose and a nitrogen atom of glutamine are joined to form a carbon-nitrogen bond. As such BPLER cells were incubated with ^{13}C-C2 glucose and ^{15}N-glutamine, and a HCN experiment was recorded to detect metabolites resonances that contain a hydrogen, connected to a carbon, that is also connected to a nitrogen atom (HCN). In this experiment, as shown in Figure 8B BPLER cells contain a differentially expressed resonance at the same position as would be expected in the NANA standard. Lastly, mass spectrometry experiments shown in Figure 8C, confirmed NANA is approximately 7-fold higher in BPLER tumor initiating cells by performing multiple reaction monitoring LC/MS using electrospray in the negative mode. The reported values are the area under the curve for NANA expression in each cell line.

Using results of differentially expressed metabolites to develop diagnostics: By following glucose flux within BPLER cells (i.e. subtracting background ^{13}C and tracing specific breakdown of glucose), the tumor initiating cells were observed to divert part of their glucose metabolism to NANA production. NANA is 9-carbon sugar that is often incorporating onto the cell surface of glycoproteins. Previous reports identified that wheat-germ agglutinin (WGA) has a strong affinity for NANA-modified proteins. Using rhodamine labeled WGA, we preformed immune-fluorescent microscopy shown in Figure 9 and observed BPLER cells have increased NANA expression on their cell surface. Thus, NANA itself represents a novel diagnostic to specifically identify breast tumor initiating cells, and WGA, and similar molecules that specifically recognize NANA and NANA modified molecules could provide new tools to detect & isolate tumor initiating cells or those with increased malignant potential.

Example 6: Identification of a target utilizing the differential NMR data

NANA is a sugar that is often incorporated onto cell surface proteins. Shown in Figure 10 NANA is derived from glucose by about 15 distinct enzymes including key enzymes NANS and CMAS. Using CelTiterGlo, a well-known assay for cell viability, it was observed that knockdown of NANS or CMAS had little to no effect on cell viability or proliferation of the cells (Figure 11), whereas knockdown of PLK1 enzyme lead to total cell death.

However, NANA is incorporated on the cell surface of several proteins involved in cell adhesion, and loss of NANA was suspected to affect cell motility. Using a cell migration assay, cells were cultured in a dual-chamber containing small pores at the bottom of the top chamber, malignant cells (especially those with metastatic potential) are able to migrate through the pores and form colonies. As expected, shown in Figure 12 BPLER have an increased migration rate as compared to HMLER cells due to their increased turmorigenic properties. However, the knockdown of NANS or CMAS completely abolished the cells' ability to migrate to the lower chamber, while cells transfected with control siRNAs maintained normal migration. To confirm NANA expression directly influences motility, a rescue experiment was performed in which cells transfected with siRNAs against NANS or CMAS were supplemented NANA. In the presence of NANA they were able to partially restore the migration phenotype. These results of several migration studies are quantified in Figure 12B. These results suggest NANA expression is crucial for cell migration and could be important for metastasis. Monitoring NANA expression could predict the metastatic potential of a cell. In addition both NANS/CMAS are novel key targets to manipulate migration and metastasis.

Example 7: CMAS increase cell migration

As mentioned, the knockdown of NANS and CMAS, key enzymes used to generate and attach NANA to proteins, had no effect on cell proliferation but greatly reduced the ability of BPLER cells to migrate (shown in Figure 12). While the mRNA level of CMAS and NANS was equivalent in HMLER and BPLER cells, we observed the protein expression of CMAS is dramatically over expressed in BPLER tumor-initiating cells (Figure 13). To determine if CMAS expression contributes to the malignant phenotype of tumor initiating

BPLER cells, non-aggressive HMLER cells were transfected with a plasmid to force the expression of CMAS. Using the previously described cell migration assay, as shown in Figure 14A HMLER cells with enhanced CMAS expression dramatically increased their migration potential by several fold compared to the control. The reciprocal experiment was also performed and stable CMAS knockdown BPLER cells (BPLER-shCMAS1) were created. As expected, shown in Figure 14B these cells were not able to form colonies in the same migration assay. These results suggest CMAS protein expression is pivotal in cell migration and/or metastasis. To date there have been no references to the role of either NANS or CMAS in cancer. This could be in part due to most techniques relying on sequencing and/or microarray experiments that only examine mRNA levels. This highlights the strength of our method. By probing metabolites, NANA was identified as being expressed significantly higher in breast tumor-initiating BPLER cells. By subsequently probing the enzymes in its biosynthetic pathway, a novel, potent target that could be important for tumorigenicity was discovered.

Example 8: CMAS impacts tumor formation initiation

To determine how loss of CMAS/NANA expression effects the tumor initiation *in vivo* we performed the experiment outlined in Figure 15A. As described above stable CMAS knockdown BPLER cells (BPLER-shCMAS1) were created that do not express CMAS protein. In our experiment 500,000 BPLER cells over expressing an empty vector (control) and 500,000 BPLER-shCMAS1 cells (which do not express CMAS) were injected into the mammary fat pad of NOD/SCID mice. The goal was to analyze differences in tumor size and the number of metastasis between each group. Every three days for 45 days the mice were examined for palpable tumors, and if detected the tumor height, length and width were directly measured to calculate tumor volume. As shown in Figure 15B, within 45days, 4/4 of the control mice had developed large primary tumors. Amazingly, none (0/5) of BPLER-shCMAS1 mice had any palpable tumors. Indeed after 90days the mice injected with BPLER-shCMAS1 cells remained tumor free. Taken together the *in vitro* and *in vivo* data suggest that CMAS is a completely novel and bona fide therapeutic target for cancer.

Example 9: Using the NMR Data to Identify and Design Therapeutic Agents for Breast Cancer Therapy

Enzymes such as CMAS are ideal candidates for small molecule drug inhibition. The enzyme mechanism of CMAS is Figure 16A, where CMAS activates the hydroxy group of NANA in a divalent cation dependent manner, so that it can subsequently attack the alpha-phosphate of an incoming cytidine triphosphate (CTP) molecule to form a cytidine monophosphate-NANA (CMP-NANA) intermediate. Using the NANA scaffold, a substrate-based analog replacing the hydroxyl group with a fluorine was designed and synthesized (Figures 16B). This substitution should theoretically maintain the ability of NANA to bind CMAS however, prevent the enzymatic reaction. In the presence of the CMAS inhibitor, using the previously described cell migration assay, it was shown that BPLER cells are no longer able to migrate (Figure 16C). Using the NANA scaffold, we were able to rapidly design and synthesize a substrate based inhibitor that has effect in cell lines.

The F-NANA derivative synthesized had a slight chemical likeness to the FDA approved drugs Relenza and Tamiflu (Figure 17). Relenza and Tamiflu are both designed to inhibit the influenza enzyme neuraminidase. Neuraminidase specifically cleaves NANA molecules on the cell-surface to facilitate viral entry into the cell. Neuraminidase and CMAS share NANA as a substrate, and hence these known influenza therapeutics were suspected to be inhibitors of CMAS. As shown in Figure 17b, Relenza treatment of BPLER cells blocked cell migration. Both Relenza and Tamiflu are already FDA approved, marketed therapeutics, and pending positive results in mouse models, a rapid entry to clinical trial for cancer indications.

Neuraminidase itself is known to remove NANA from the cell surface. We suspected neuraminidase could be used to remove NANA from the surface of malignant cells and just like siRNAs against CMAS exert a similar effect on migration and tumor initiation. Pre-incubation of BPLER cells with active neuraminidase enzyme diminished NANA expression as determined by rhodamine labeled wheat germ agglutinin (WGA) microscopy (Figure 18A). In addition, BPLER cells treated with neuraminidase were no longer able to migrate in the migration assay (Figure 18B). Neuraminidase and flu-like molecules, including empty virions, may represent an innovative way to both detect tumor-initiating cells (influenza

virions have a high affinity for NANA) and inhibit tumor initiation and metastasis by removing NANA from tumor populations.

OTHER EMBODIMENTS

It is to be understood that while the inventions have been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the inventions, which are defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

What is claimed is:

1. A method for monitoring metabolism of a substrate within a given type of cell in a sample, the method comprising:

a. culturing a given type of cell of a first sample with a substrate for a sufficient period of time to allow metabolic breakdown of the substrate into substrate metabolites, wherein at least a portion of the substrate is optionally labeled with a nuclear magnetic resonance (NMR) stable isotope;

b. harvesting the substrate metabolites from the cells of step (a) to obtain a second sample of substrate metabolites; and

c. performing multi-dimensional NMR on the second sample of step (b) to determine a resonance spectrum of the metabolized substrate, wherein the resonance spectrum represents the metabolites of the substrate, and wherein the multi-dimensional NMR comprises any one of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

2. The method of claim 1, wherein the multi-dimensional NMR comprises any two, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

3. The method of claim 1, wherein the multi-dimensional NMR comprises any three, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

4. The method of claim 1, wherein the multi-dimensional NMR comprises all four of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

5. The method of claim 1, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises any two, in any combination, of the following

techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

6. The method of claim 1, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises non-uniform sampling and data extension for enhanced dynamic range data reconstruction.

7. The method of claim 1, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

8. The method of claim 1, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises all four of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

9. The method of claim 1, wherein the substrate metabolites that are present in the sample are not purified away from the other molecules in the sample.

10. The method of any one of claims 1 to 9, wherein the substrate concentration within the population of cells is reduced for a period of time prior to loading the cells with the NMR-labeled substrate.

11. The method of any one of claims 1 to 9, wherein the resonances of the metabolites of the labeled substrate are determined using NMR pulse programs or filtering techniques, or both, customized to the substrate.

12. The method of any one of claims 1 to 9, wherein the number of cells within the population of cells is less than 2×10^6.

13. The method of any one of claims 1 to 9, wherein the population of cells is a primary population of cells.

14. A method for identifying differentially expressed substrate metabolites between a first population of cells and a second population of cells, the method comprising:
    a. optionally loading a first and a second population of cells with a nuclear magnetic resonance (NMR) stable isotope-labeled substrate;
    b. culturing the first and the second population of cells of step (a) for a sufficient period of time to allow metabolic breakdown of the substrate into substrate metabolites;
    c. harvesting the substrate metabolites from the first and the second population cells of step (b) to obtain a sample of substrate metabolites from each of the first and the second cell populations;
    d. performing multi-dimensional NMR on the sample of step (c) for each of the first and the second cell populations to determine a resonance spectrum of the metabolized substrate of the first population of cells and of the second population of cells, wherein the resonance spectrum represents the metabolites of the substrate; and
    e. comparing the resonance spectrum of the first population of cells with the resonance spectrum of the second population of cells to determine which resonances are differentially expressed, wherein the differentially expressed resonances provide a resonance signature that represents differentially expressed metabolites.

15. The method of claim 14, wherein the multi-dimensional NMR comprises any one of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

16. The method of claim 14, wherein the multi-dimensional NMR comprises any two, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

17. The method of claim 14, wherein the multi-dimensional NMR comprises any three, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

18. The method of claim 14, wherein the multi-dimensional NMR comprises all four of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

19. The method of claim 14, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises any two, in any combination, of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

20. The method of claim 14, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises non-uniform sampling and data extension for enhanced dynamic range data reconstruction.

21. The method of claim 14, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

22. The method of claim 14, wherein the substrate is labeled with a stable isotope and the multi-dimensional NMR comprises all four of the following techniques: spectral width folding, random phase sampling, non-uniform sampling, and data extension for enhanced dynamic range data reconstruction.

23. The method of claim 14, wherein the substrate metabolites that are present in the sample are not purified away from the other molecules in the sample.

24. The method of any one of claims 14-23, wherein substrate concentration within each population of cells is reduced for a period of time prior to loading the cells with the NMR stable isotope-labeled substrate.

25. The method of any one of claims 14-23, wherein the resonances of the metabolites of the labeled substrate are determined using any one or more of custom NMR adiabatic pulse programs, custom acquisition techniques, and custom NMR analysis and filtering programs.

26. The method of any one of claims 14-23, wherein the number of cells within each population of cells is less than 2×10^6.

27. The method of any one of claims 14-23, wherein the first cell population and second cell population are populations of primary cells.

28. The method of any one of claims 14-23, wherein the population of cells in each of the first cell population and second cell population is a heterogeneous population of cells, or is a homogeneous population of cells.

29. The method of any one of claims 14-23, further comprising comparing the resonance signature of step (e) with a database of known resonance signatures to determine the molecular structure that the resonance signature represents, and thereby determine the substrate metabolites that are differentially expressed between the first and the second population of cells.

30. The method of any one of claims 14-23, further comprising identifying a biosynthetic pathway involved in generation of the substrate metabolites and identifying proteins/enzymes of the pathway that may be targeted to modulate the differential expression of the metabolite, to thereby modulate the phenotype of the cells.

31. The method of any one of claims 14-23, wherein the first population of cells and the second population of cells are isogenic populations.

32. The method of any one of claims 14-23, wherein the first population of cells and the second population of cells have different phenotypes.

33. The method of any one of claims 14-23, wherein the first population of cells is a control population of cells and the second population of cells has been contacted with a test compound or agent.

34. The method of any one of claims 14-23, wherein the method is used to identifying metabolic pathways that are overactive or underactive in a particular cell type.

35. The method of any one of claims 14-23, wherein the method further comprises inhibiting or overexpressing a gene in the second population of cells and the method is used to identify the metabolic consequences of over-expressing or inhibiting a gene in a cell.

36. A method for treating cancer in a subject, the method comprising administering to a subject in need thereof an inhibitor of N-acylneuraminate cytidylyltransferase (CMAS), or an inhibitor of N-acetylneuraminic acid synthase (NANS), or a molecule that decreases the expression of N-acetylneuraminic acid.

37. The method of claim 36, wherein the inhibitor is selected from the group consisting of: a small molecule, a ribonucleic acid, a deoxyribonucleic acid, a protein, a peptide, and an antibody.

38. The method of claim 36, wherein the inhibitor is an enzyme.

ABSTRACT

Methods that enable one to specifically measure the metabolic product of a particular molecule in relatively few cells, e.g. primary cells, are described. The methods involve optionally preloading cells with labeled substrate (e.g. labeled by ^{13}C, ^{15}N, or ^{31}P). The methods allow for easy identification of metabolites that are differentially generated in cells of different phenotypes. The new methods for unbiased multi-dimensional NMR screening and rapid and efficient analysis of the NMR screening identify differentially expressed metabolites in different cell or tissue types. Analysis of the differentially expressed metabolites can present unique druggable targets to which small molecule therapeutics can be designed.

Figure 1.

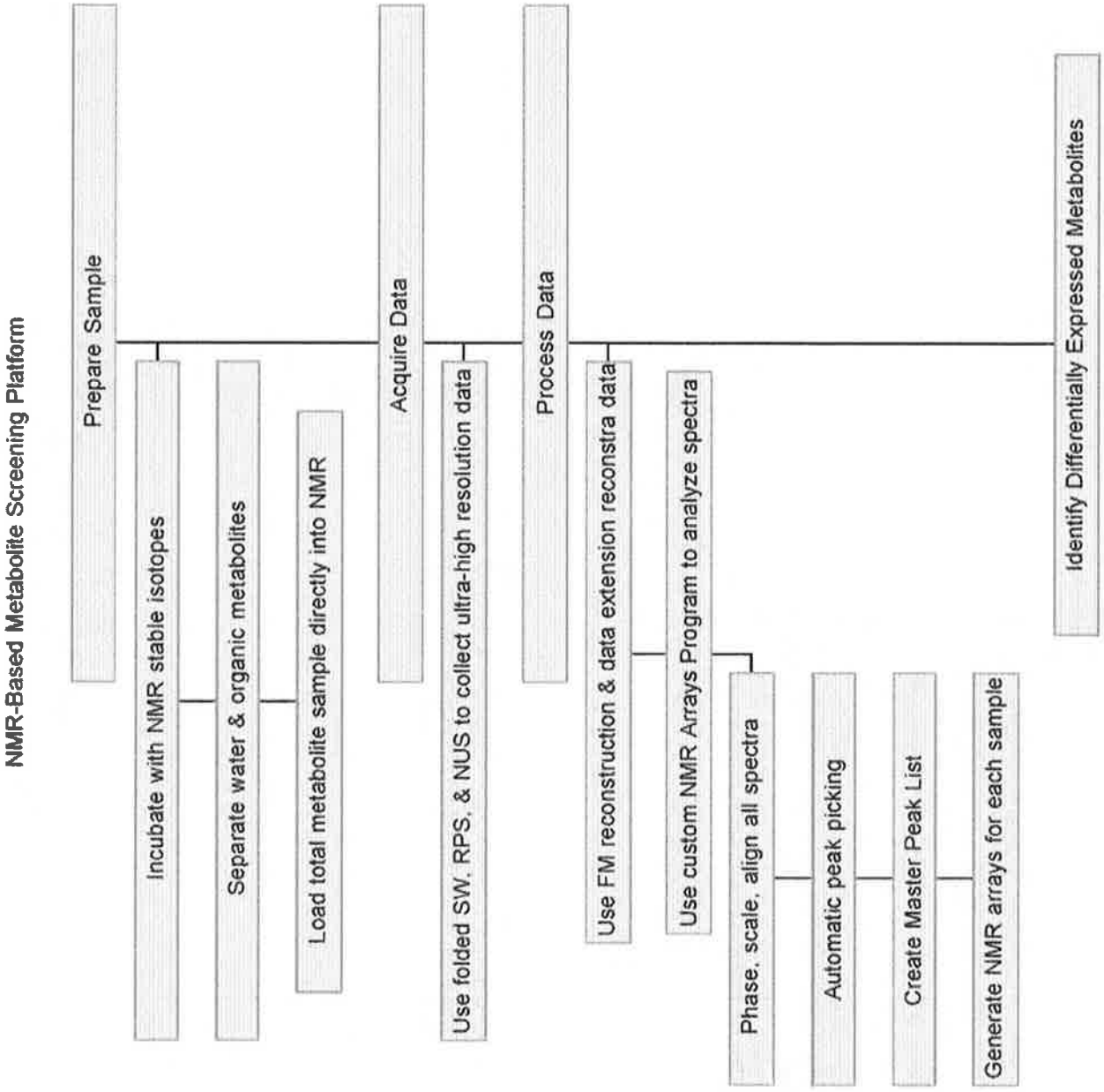
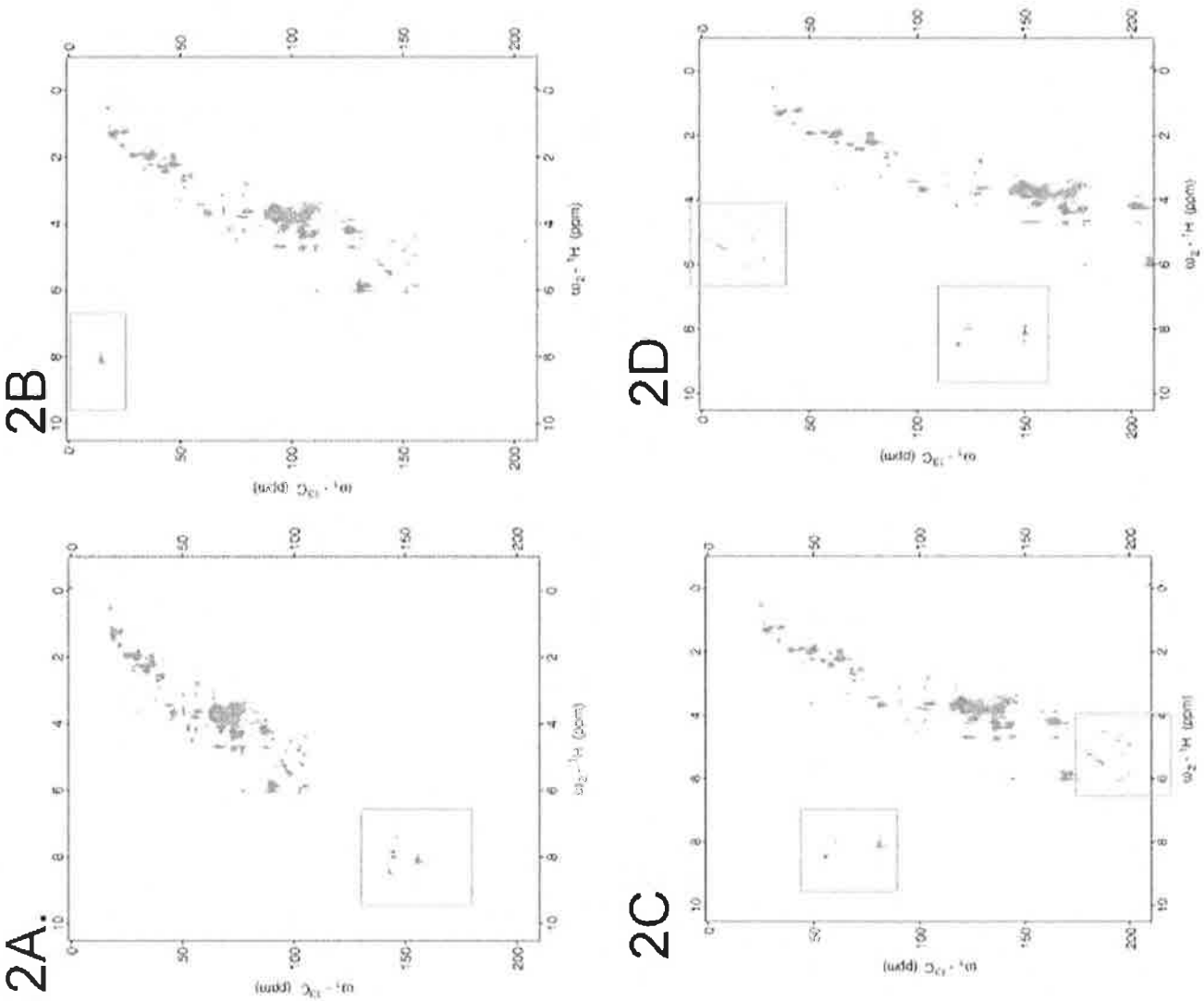

NMR-Based Metabolite Screening Platform
Prepare Sample
Incubate with NMR stable isotopes
Separate water & organic metabolites
Load total metabolite sample directly into NMR
Acquire Data
Use folded SW, RPS, & NUS to collect ultra-high resolution data
Process Data
Use FM reconstruction & data extension reconstra data
Use custom NMR Arrays Program to analyze spectra
Phase, scale, align all spectra
Automatic peak picking
Create Master Peak List
Generate NMR arrays for each sample
Identify Differentially Expressed Metabolites

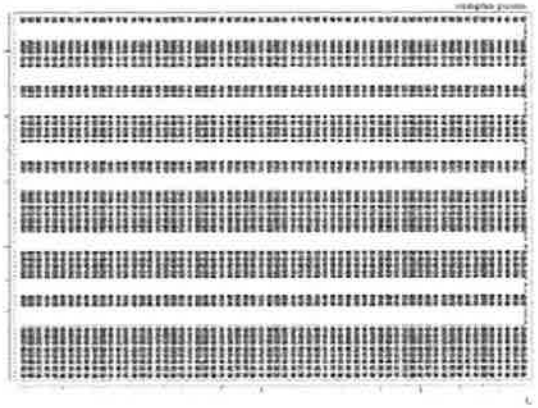
Figure 2A.



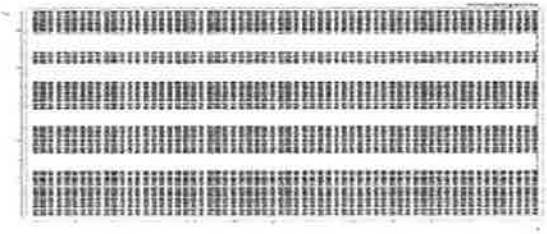
2E.

Uniform sampling of 128 indirect points requiring 5 hours experiment time with X resolution

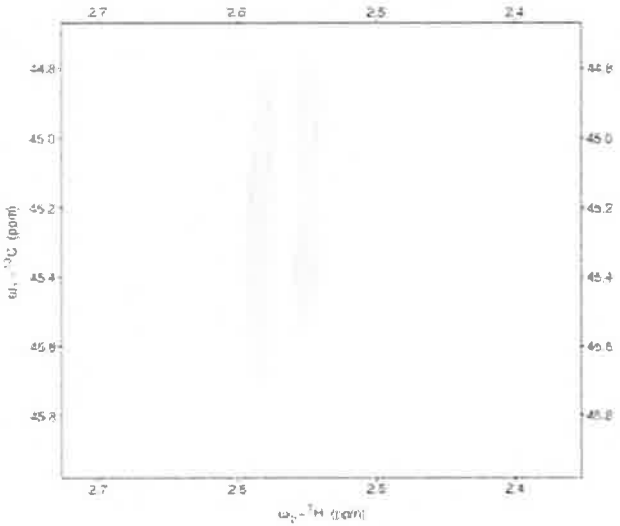
10% NUS of 1024 indirect points requiring 5 hours experiment time with 8X resolution



10% NUS of 512 indirect points requiring 2 hours experiment time 4X resolution



2F



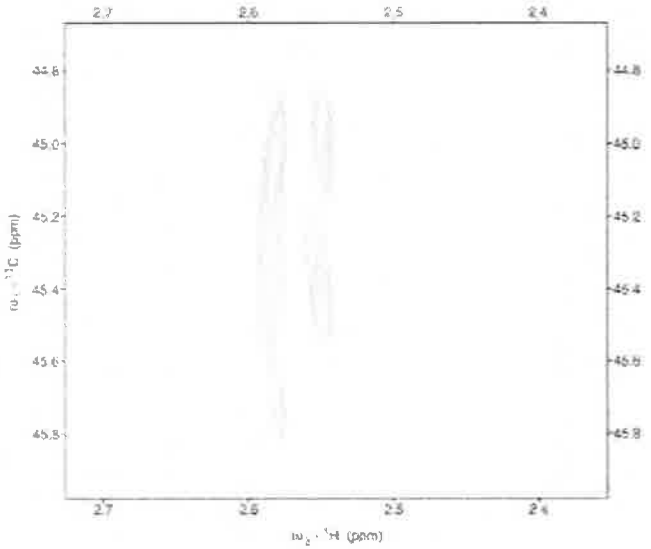
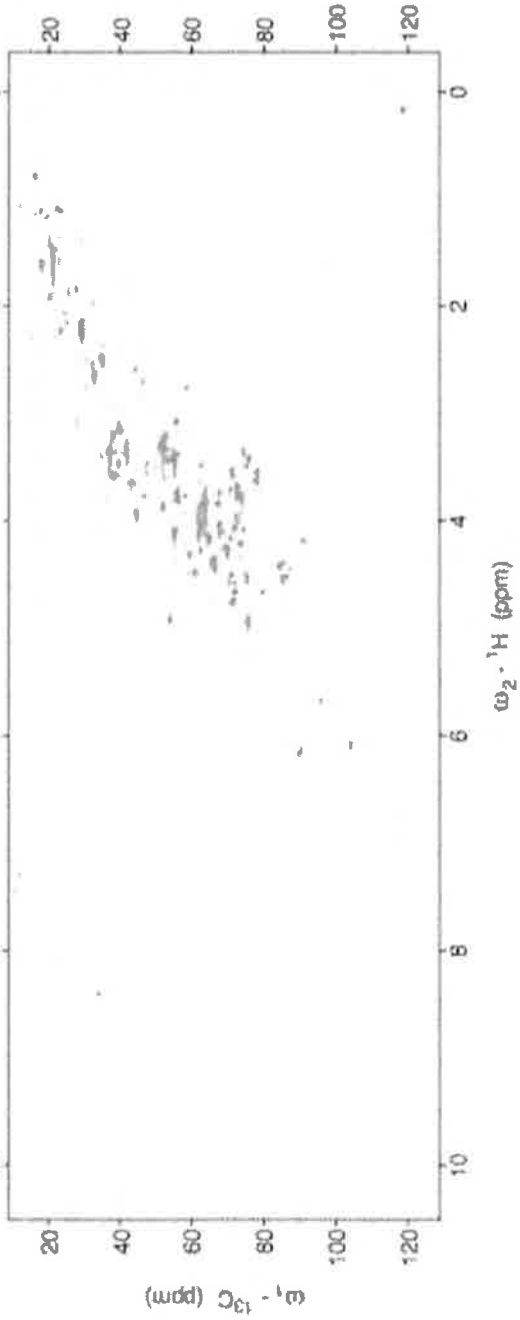


Figure 3A.



3B

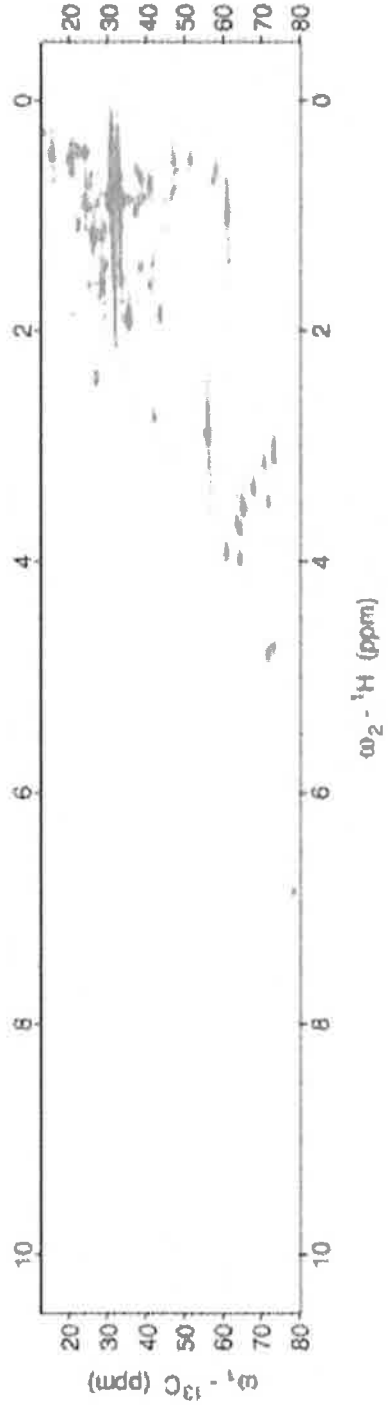


Figure 4.

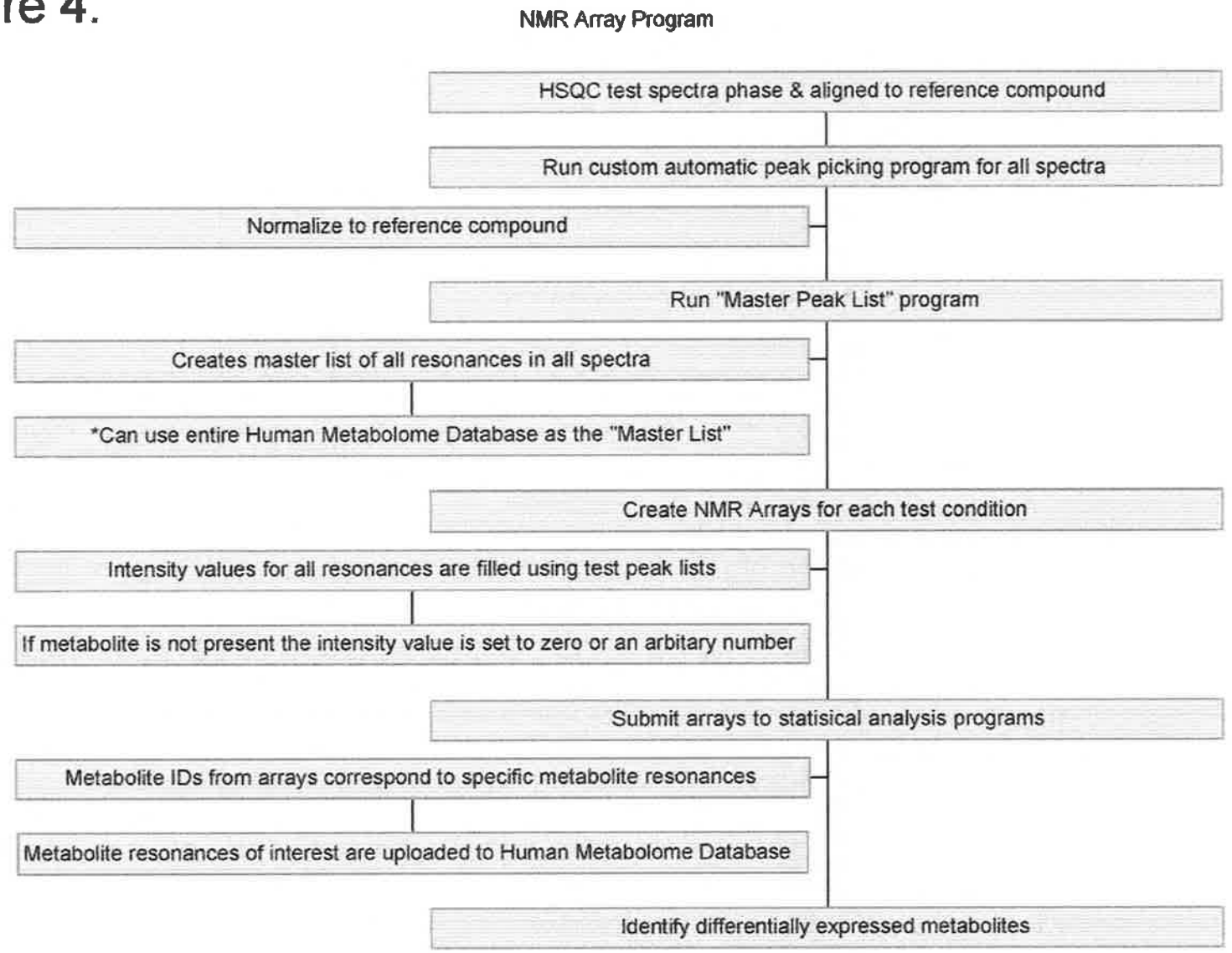

NMR Array Program
HSQC test spectra phase & aligned to reference compound
Run custom automatic peak picking program for all spectra
Normalize to reference compound
Run "Master Peak List" program
Creates master list of all resonances in all spectra
*Can use entire Human Metabolome Database as the "Master List"
Create NMR Arrays for each test condition
Intensity values for all resonances are filled using test peak lists
If metabolite is not present the intensity value is set to zero or an arbitary number
Submit arrays to statisical analysis programs
Metabolite IDs from arrays correspond to specific metabolite resonances
Metabolite resonances of interest are uploaded to Human Metabolome Database
Identify differentially expressed metabolites

Figure 5.

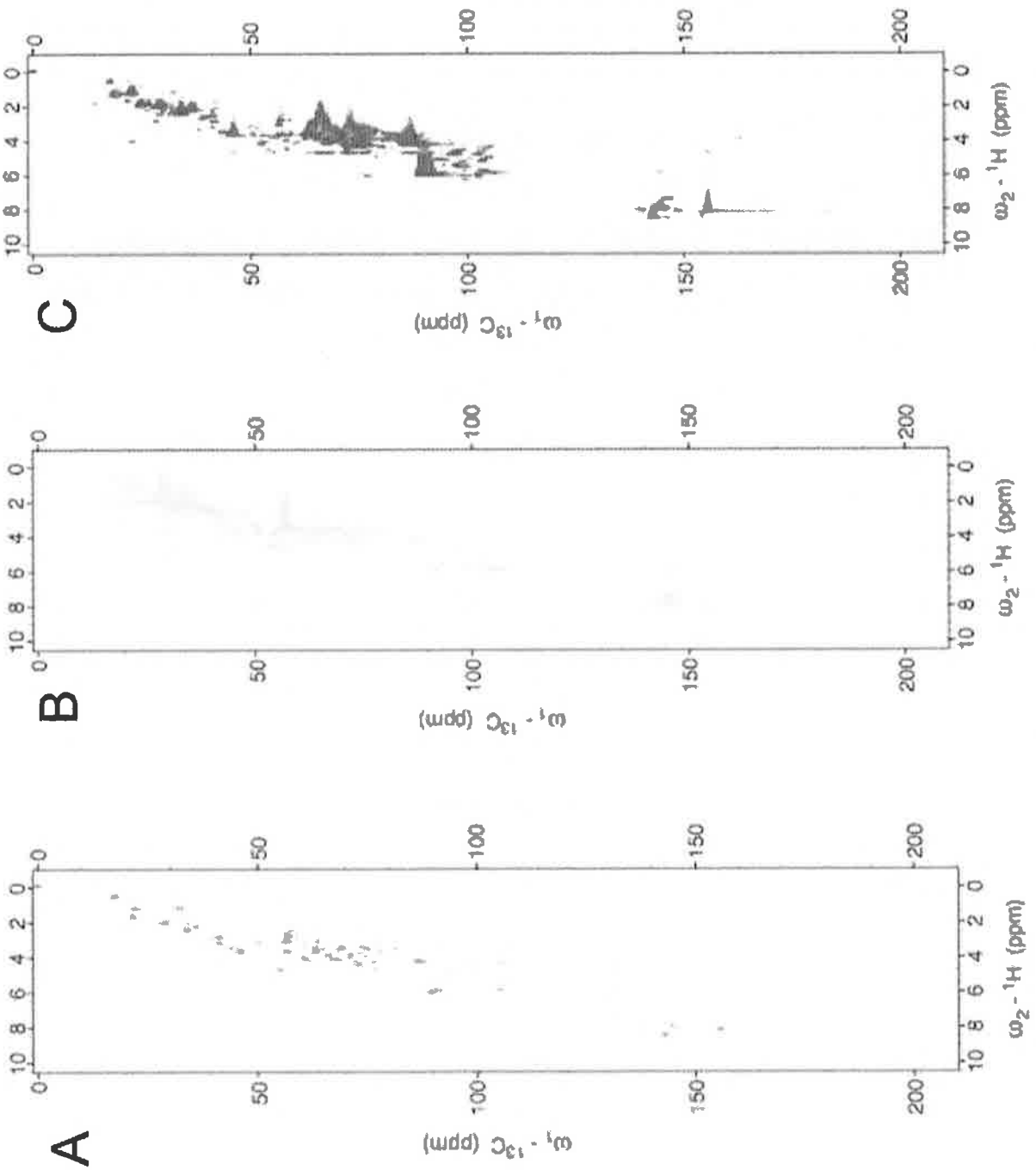


Figure 6.

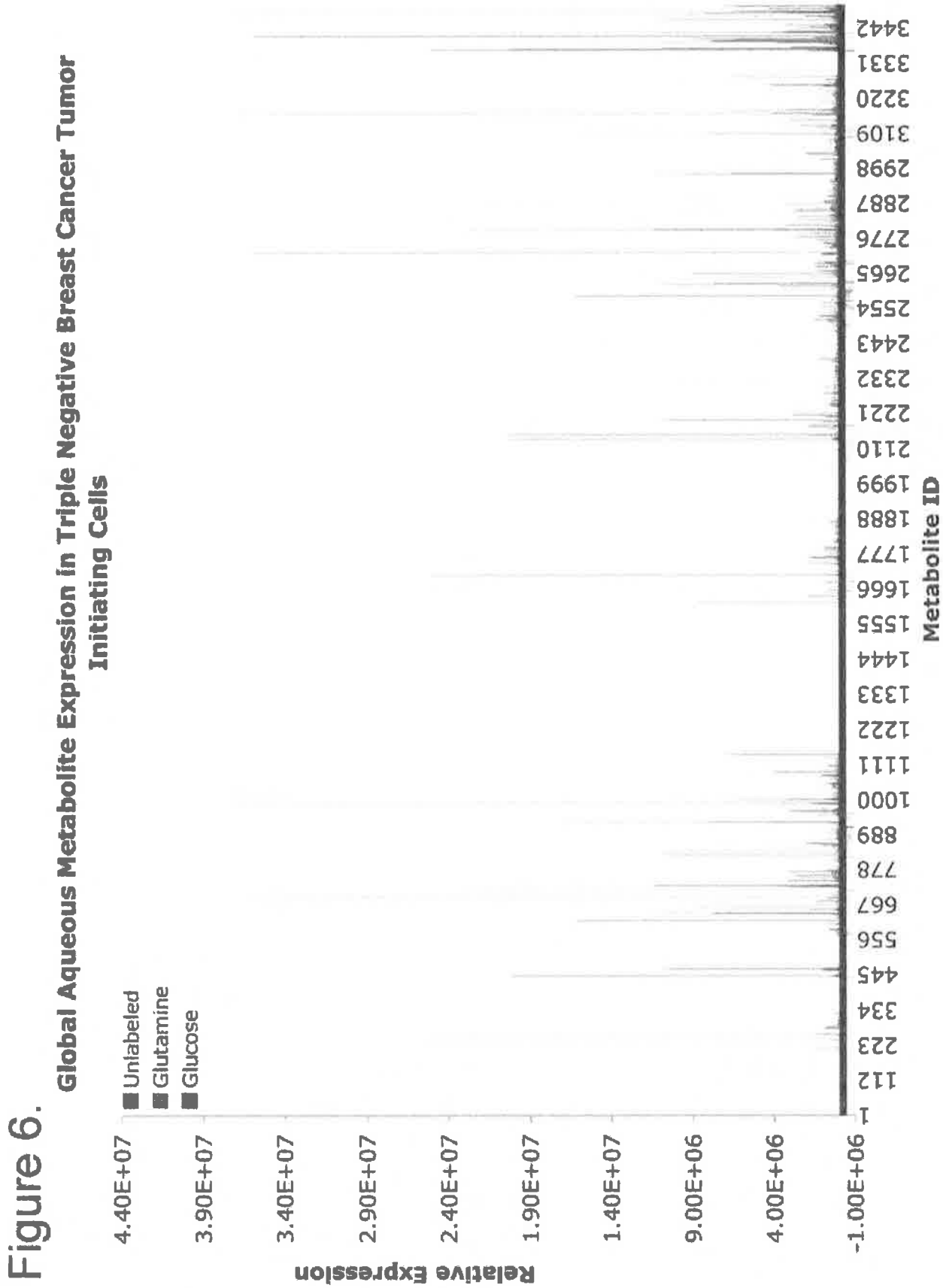

Global Aqueous Metabolite Expression in Triple Negative Breast Cancer Tumor Initiating Cells
Unlabeled
Glutamine
Glucose
Relative Expression
4.40E+07
3.90E+07
3.40E+07
2.90E+07
2.40E+07
1.90E+07
1.40E+07
9.00E+06
4.00E+06
-1.00E+06
Metabolite ID
1
112
223
334
445
556
667
778
889
1000
1111
1222
1333
1444
1555
1666
1777
1888
1999
2110
2221
2332
2443
2554
2665
2776
2887
2998
3109
3220
3331
3442

Figure 7.

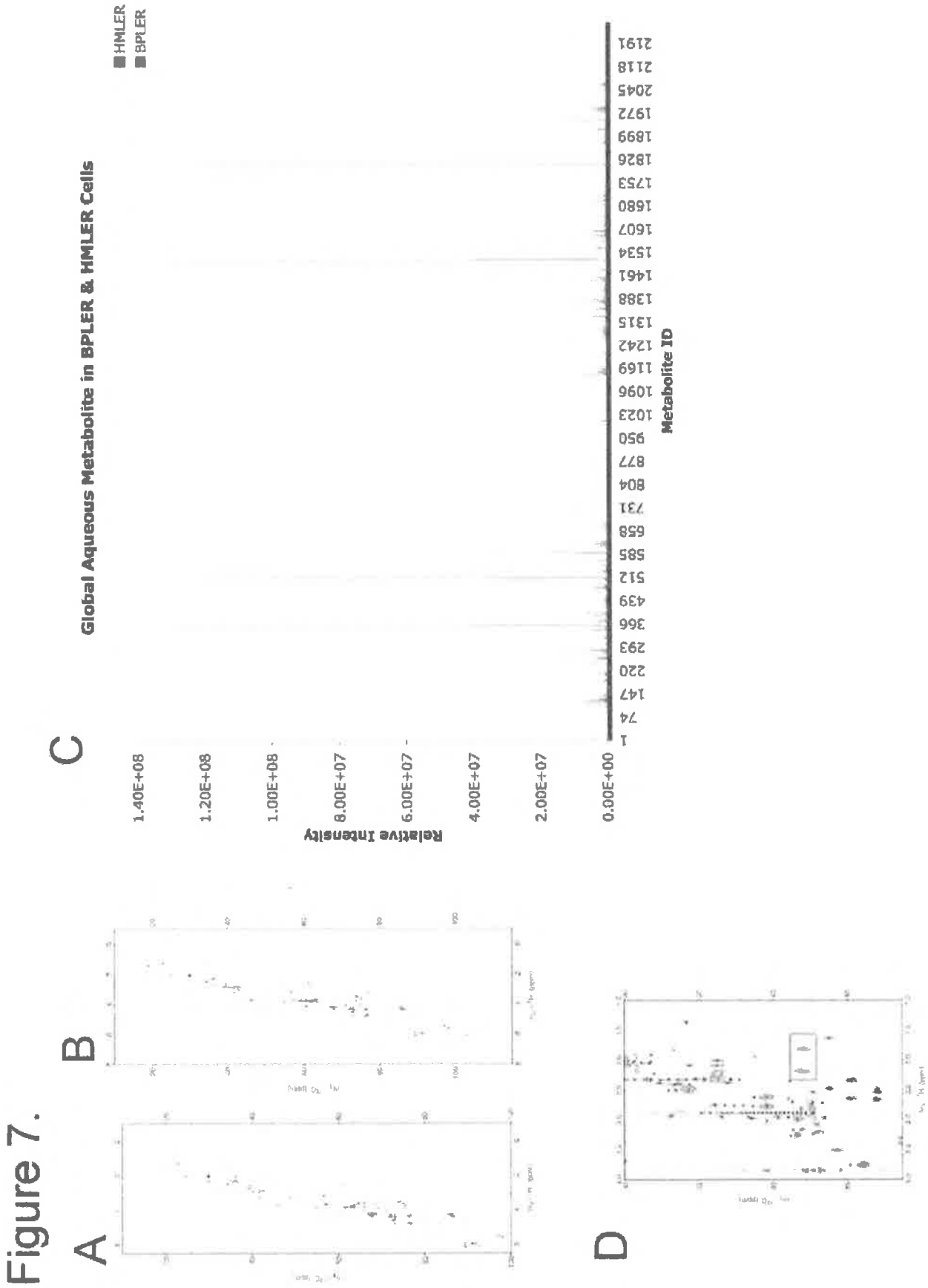
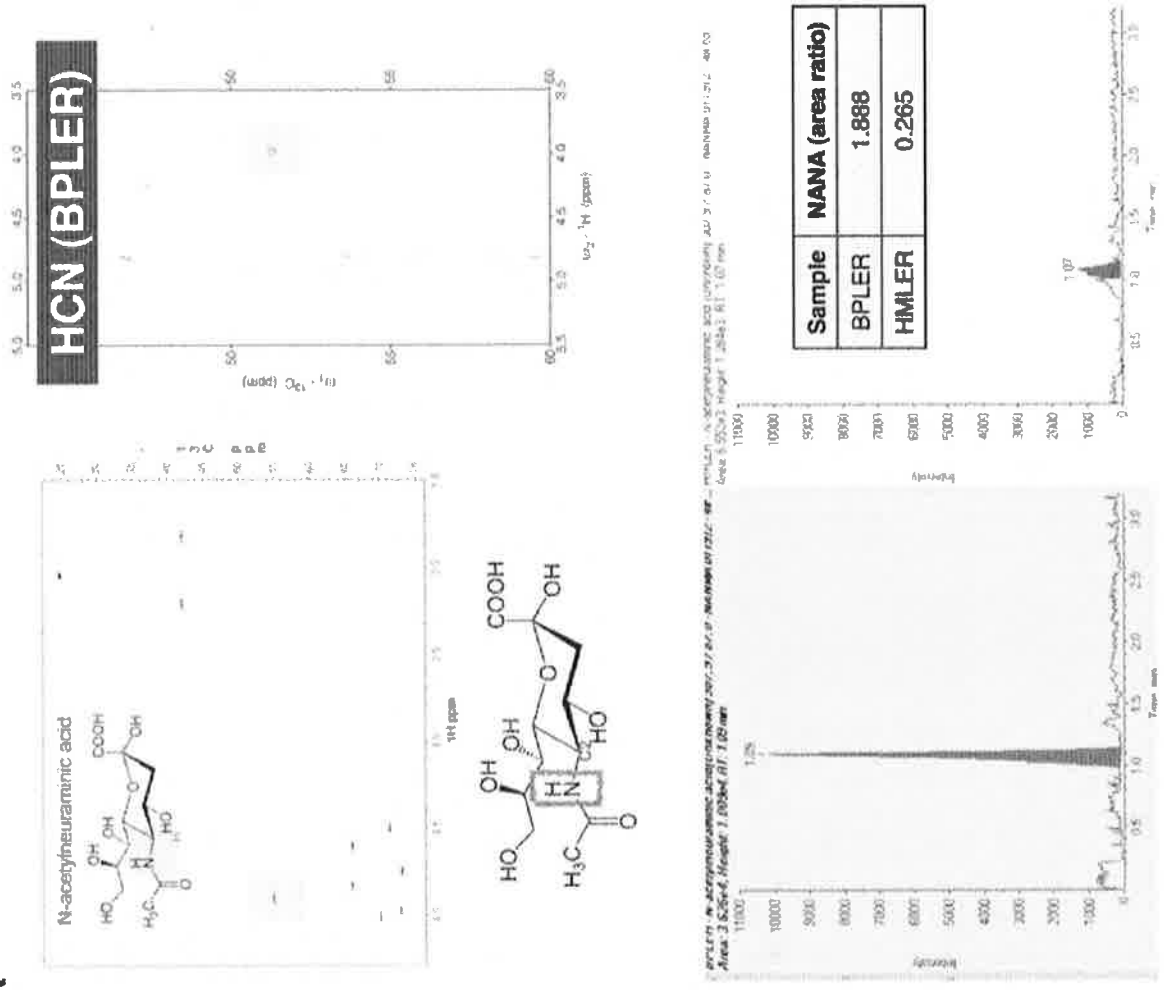

A
B
C
D
HMLER
BPLER
Global Aqueous Metabolite in BPLER & HMLER Cells
Relative Intensity
1.40E+08
1.20E+08
1.00E+08
8.00E+07
6.00E+07
4.00E+07
2.00E+07
0.00E+00
Metabolite ID
1
74
147
220
293
366
439
512
585
658
731
804
877
950
1023
1096
1169
1242
1315
1388
1461
1534
1607
1680
1753
1826
1899
1972
2045
2118
2191

Figure 8.


A
B
C
N-acetylneuraminic acid
HCN (BPLER)
Sample
NANA (area ratio)
BPLER
1.888
HMLER
0.265

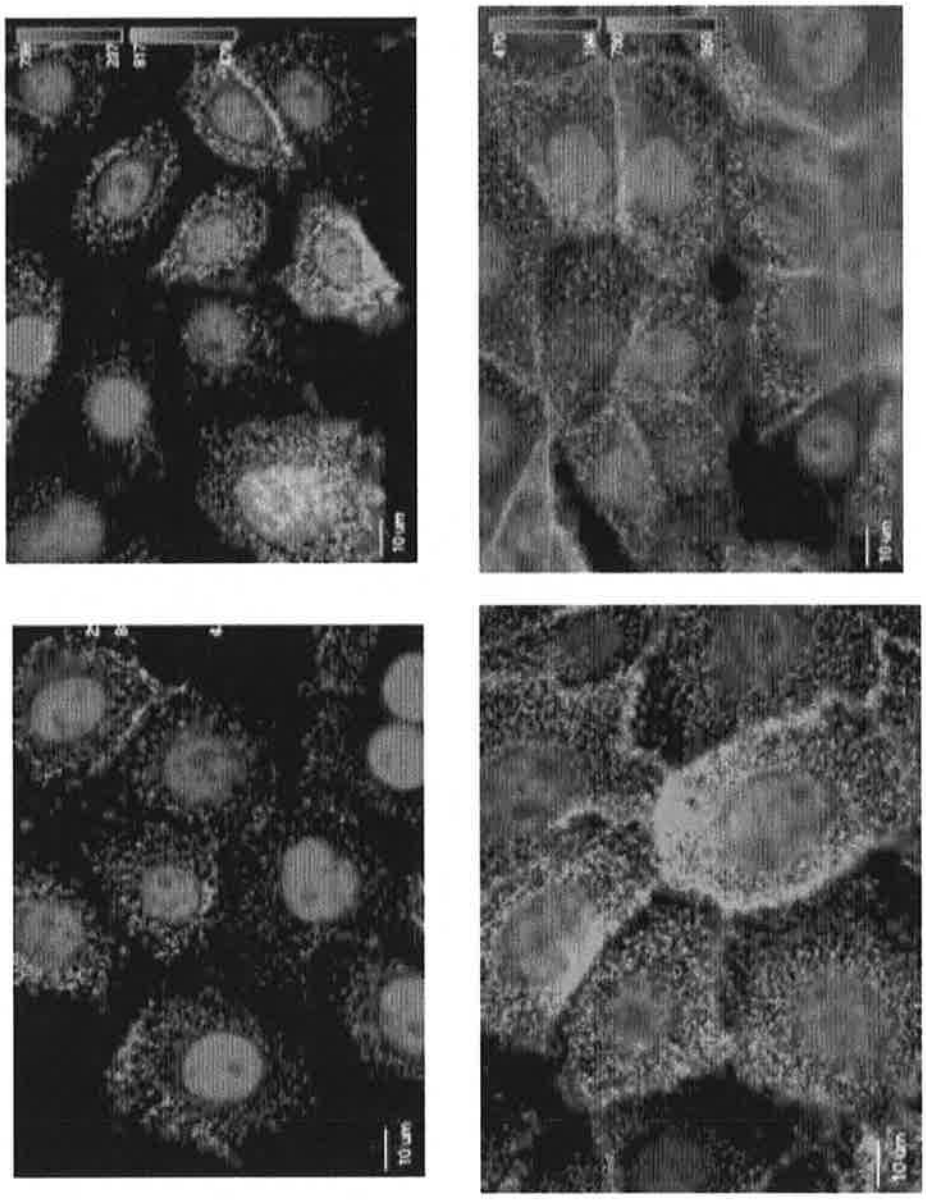


Figure 9.

Figure 10.

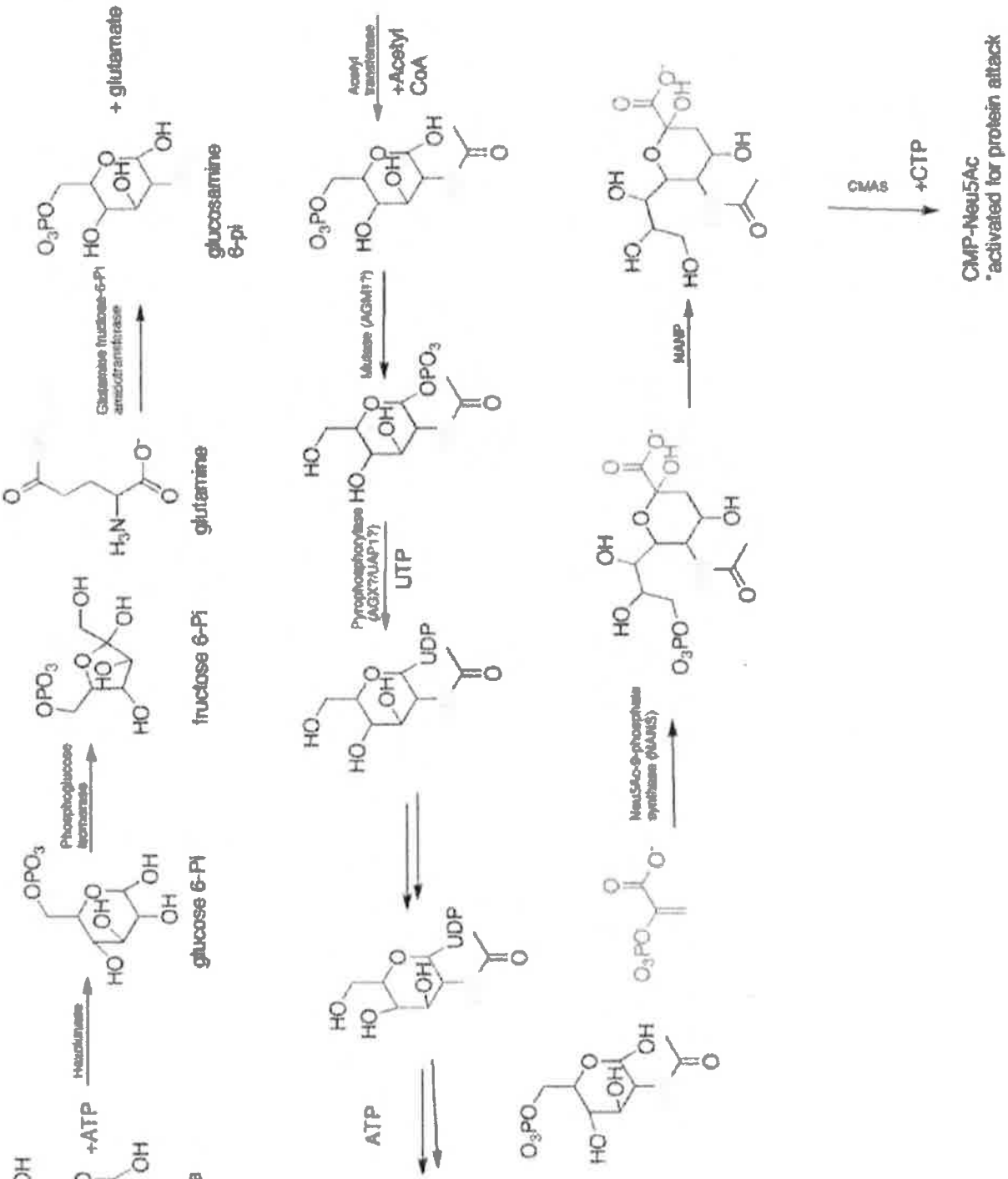
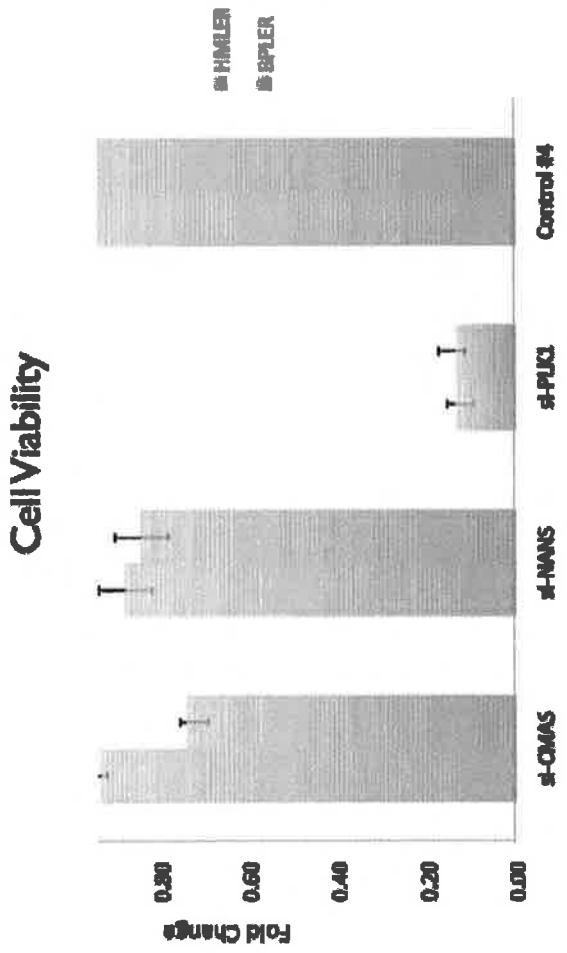

glucose
+ATP
Hexokinase
glucose 6-Pi
Phosphoglucose isomerase
fructose 6-Pi
glutamine
Glutamine fructose-6-Pi amidotransferase
+ glutamate
glucosamine 6-pi
Acetyl transferase
+Acetyl CoA
Mutase (AGM1?)
Pyrophosphorylase (AGX1?/UAP1?)
UTP
ATP
UDP
Neu5Ac-9-phosphate synthase (NANS)
NANP
CMAS
+CTP
CMP-Neu5Ac
*activated for protein attack

Figure 11.


Cell Viability
HMLER
BPLER
Fold Change
0.00
0.20
0.40
0.60
0.80
si-OMA5
si-HMN5
si-PLK1
Control #4

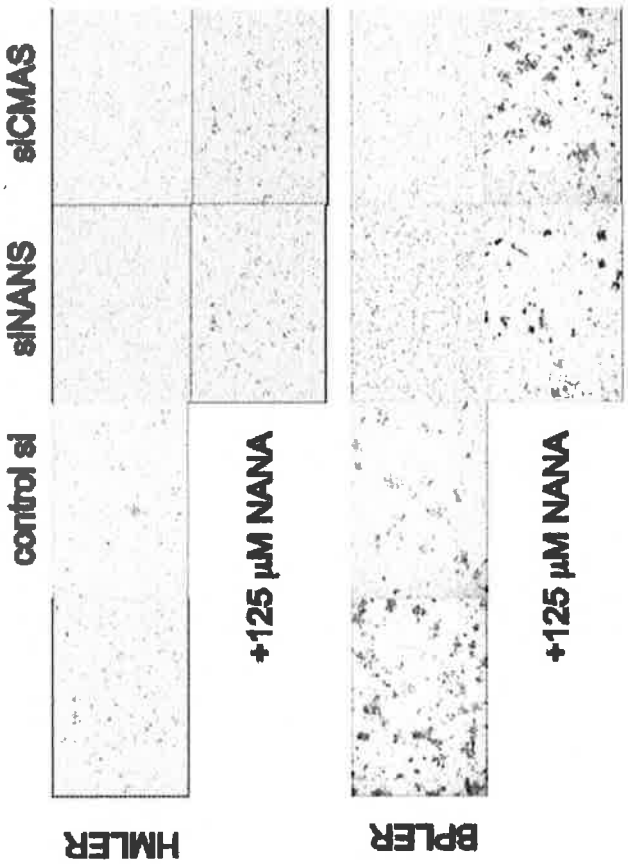


FIGURE 12A

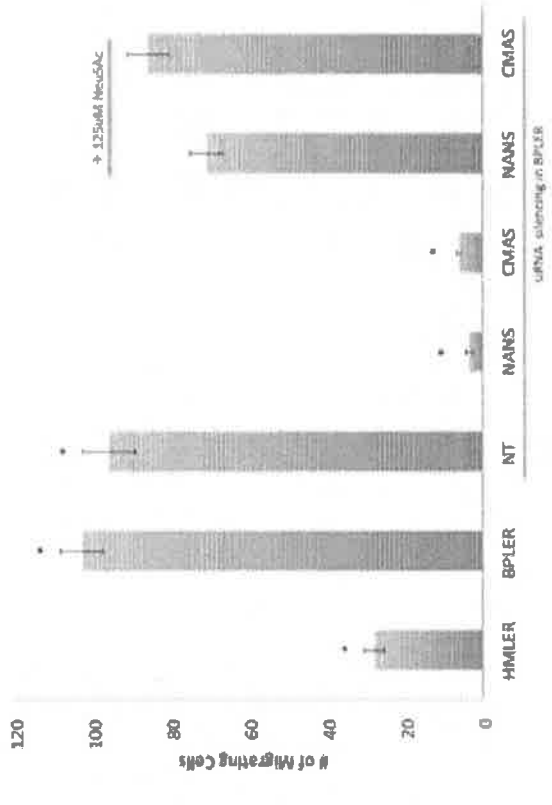


FIGURE 12B

FIGURE 12C

FIGURE 13B

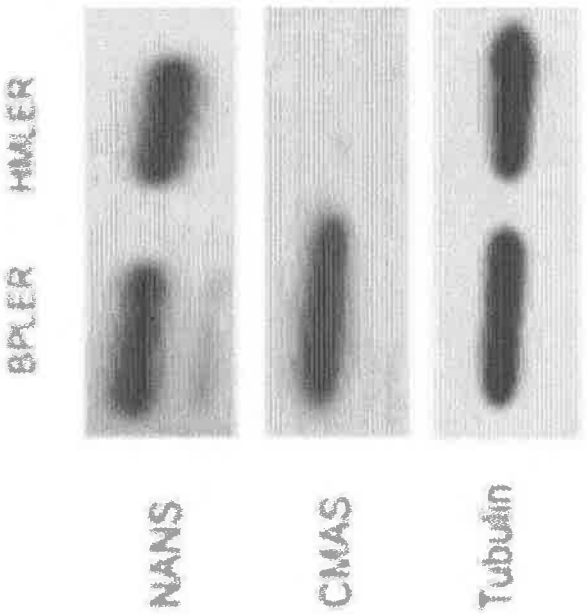
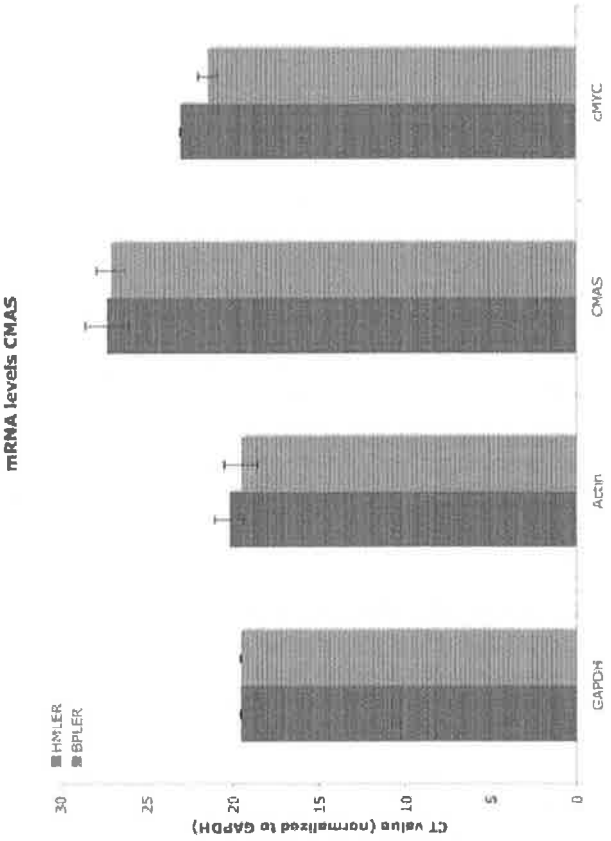


FIGURE 13A



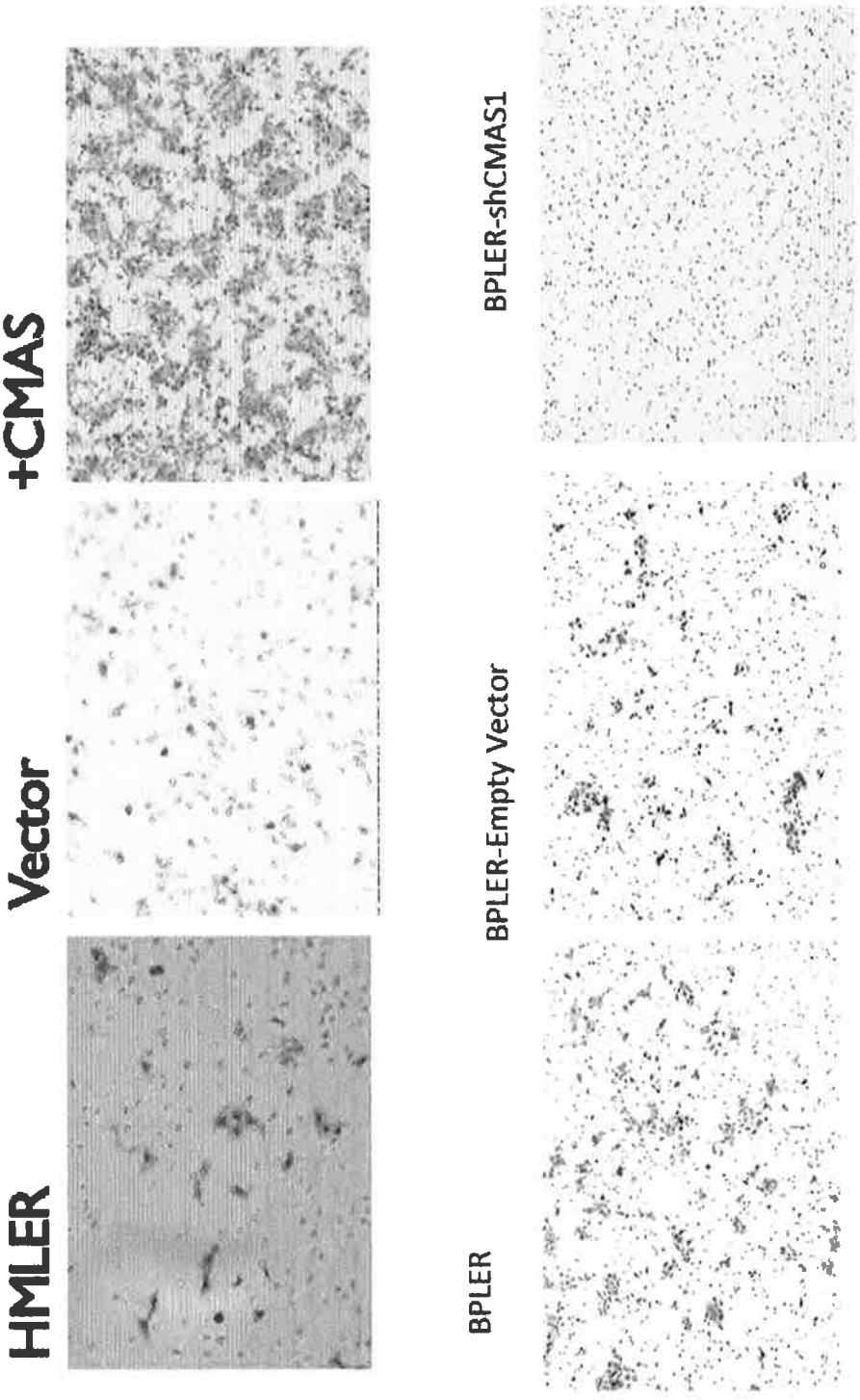
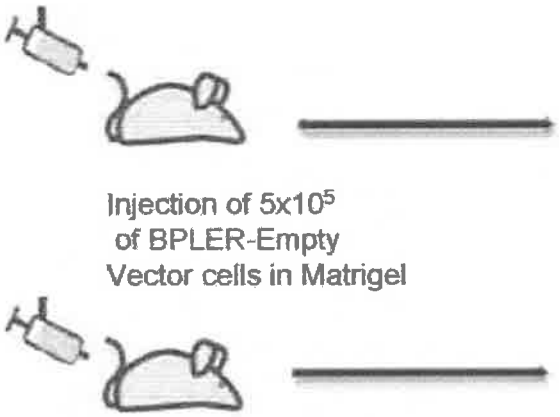


FIGURE 14A

FIGURE 14B

FIGURE 15A



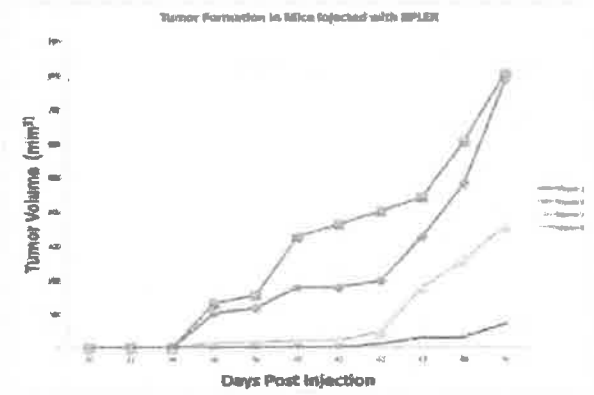
Injection of $5x10^5$ of BPLER-Empty Vector cells in Matrigel

Injection of $5x10^5$ BPLER-shCMAS1 cells in Matrigel

-Harvest breast tumors, lungs and liver

-Analyze metastasis by fluorescence and histology

FIGURE 15B



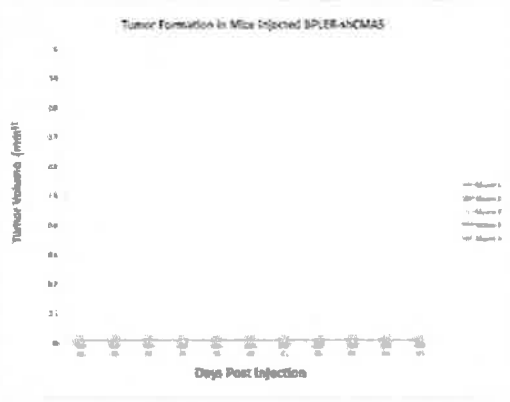


FIGURE 16A

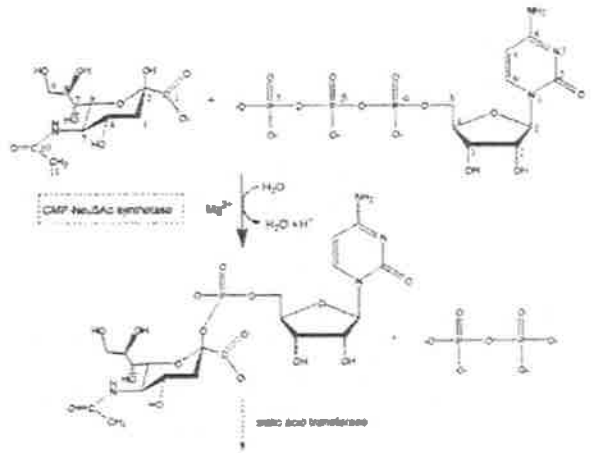


FIGURE 16B

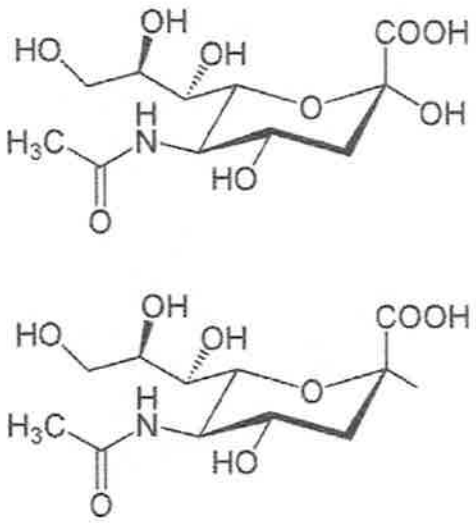


FIGURE 16C

BPLER

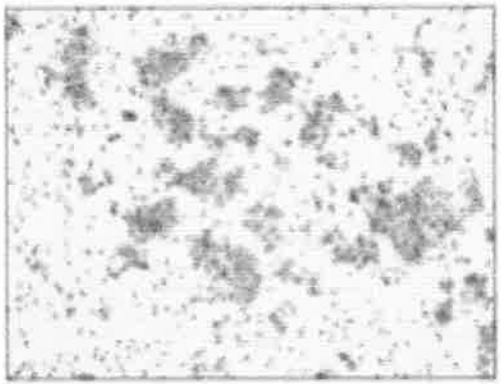
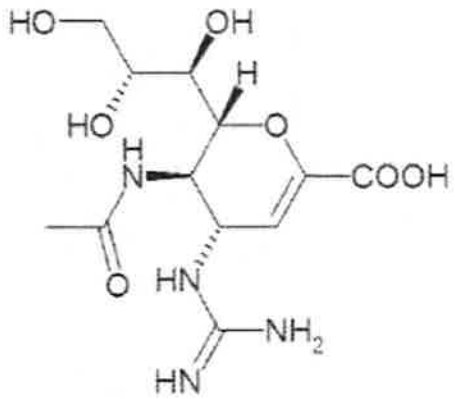


FIGURE 16D

BPLER + F-NANA

FIGURE 17A

Relenza



Tamiflu

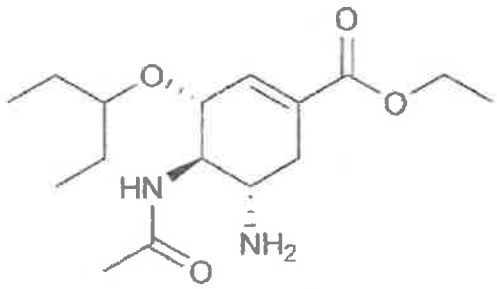
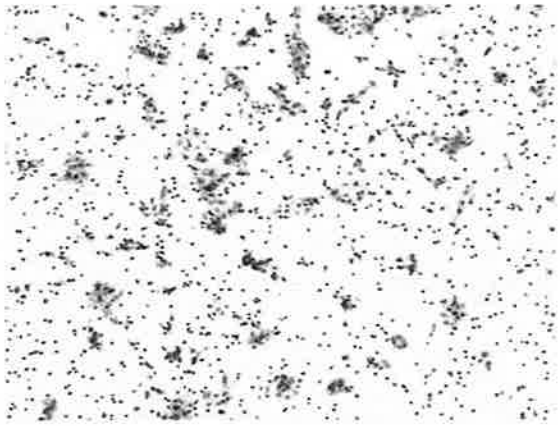


FIGURE 17C

FIGURE 17B

BPLER



BPLER- Relezna 1mg

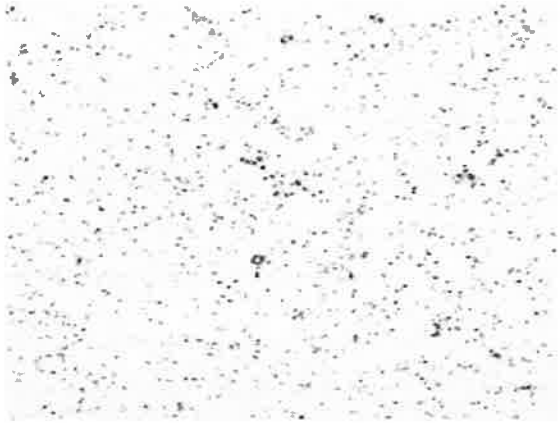


FIGURE 17D

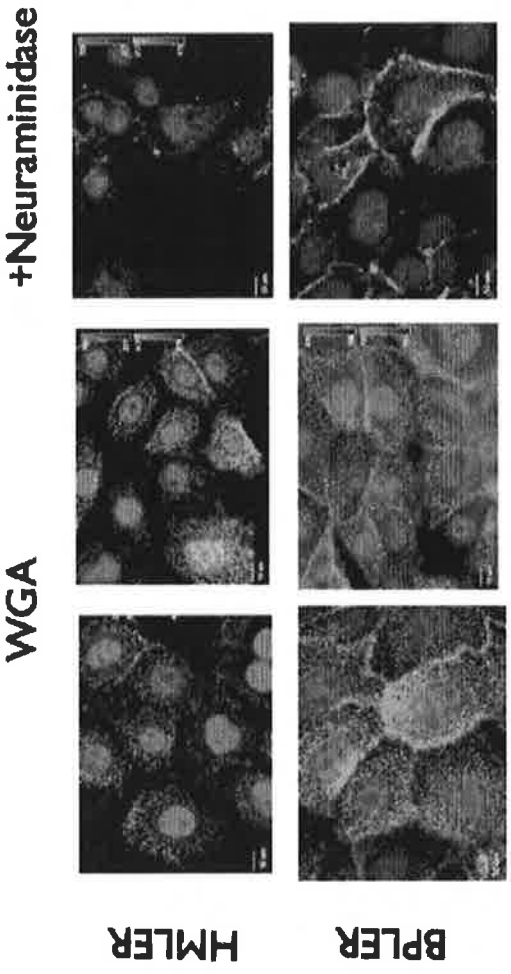


FIGURE 18A

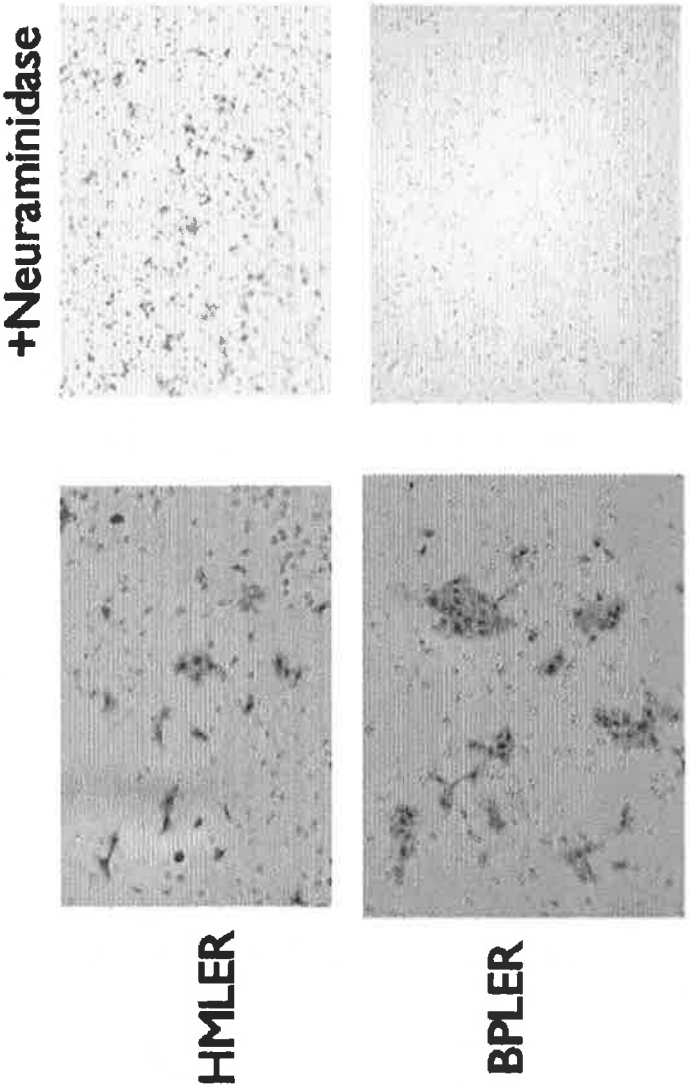


FIGURE 18B

Cell Reports
Article

The C-Terminal Domain of Eukaryotic Initiation Factor 5 Promotes Start Codon Recognition by Its Dynamic Interplay with eIF1 and eIF2β

Rafael E. Luna,[1,8] Haribabu Arthanari,[1,8] Hiroyuki Hiraishi,[2] Jagpreet Nanda,[3] Pilar Martin-Marcos,[4] Michelle A. Markus,[1,9] Barak Akabayov,[1] Alexander G. Milbradt,[1] Lunet E. Luna,[1,5,10] Hee-Chan Seo,[6] Sven G. Hyberts,[1] Amr Fahmy,[1] Mikhail Reibarkh,[1,11] David Miles,[2] Patrick R. Hagner,[1] Elizabeth M. O'Day,[1] Tingfang Yi,[1] Assen Marintchev,[7] Alan G. Hinnebusch,[4] Jon R. Lorsch,[3] Katsura Asano,[2] and Gerhard Wagner[1,*]

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA
[2]Molecular, Cellular and Developmental Biology Program, Division of Biology, Kansas State University, Manhattan, KS 66506, USA
[3]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[4]Laboratory of Gene Regulation and Development, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA
[5]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Department of Molecular Biology, University of Bergen, Bergen 5020, Norway
[7]Department of Physiology and Biophysics, Boston University School of Medicine, Boston, MA 02118, USA
[8]These authors contributed equally to this work
[9]Present address: R&D Division, Bruker BioSpin Corporation, Billerica, MA 01821, USA
[10]Present address: Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA 94720, USA
[11]Present address: Merck & Co. Inc., Rahway, NJ 07065, USA
*Correspondence: gerhard_wagner@hms.harvard.edu
DOI 10.1016/j.celrep.2012.04.007

SUMMARY

Recognition of the proper start codon on mRNAs is essential for protein synthesis, which requires scanning and involves eukaryotic initiation factors (eIFs) eIF1, eIF1A, eIF2, and eIF5. The carboxyl terminal domain (CTD) of eIF5 stimulates 43S preinitiation complex (PIC) assembly; however, its precise role in scanning and start codon selection has remained unknown. Using nuclear magnetic resonance (NMR) spectroscopy, we identified the binding sites of eIF1 and eIF2β on eIF5-CTD and found that they partially overlapped. Mutating select eIF5 residues in the common interface specifically disrupts interaction with both factors. Genetic and biochemical evidence indicates that these eIF5-CTD mutations impair start codon recognition and impede eIF1 release from the PIC by abrogating eIF5-CTD binding to eIF2β. This study provides mechanistic insight into the role of eIF5-CTD's dynamic interplay with eIF1 and eIF2β in switching PICs from an open to a closed state at start codons.

INTRODUCTION

To achieve accurate selection of the AUG start codon, the scanning preinitiation complex (PIC) is thought to exist in equilibrium between two conformations, mediated by initiation factors: open, scanning competent, and closed, scanning incompetent (Pestova and Kolupaeva, 2002). Recently, a crystal structure of the *Tetrahymena thermophila* 40S subunit along with eukaryotic initiation factor (eIF) 1 was determined, which is strategically located near the P-site (Rabl et al., 2011) consistent with previous hydroxyl-radical footprinting studies (Lomakin et al., 2003). In this crystal structure, the head of the 40S in different space groups has been found in different conformations relative to the body. This may indeed be related to the two inferred states; however, the detailed mechanism of start codon recognition caused by the dynamic interplay of initiation factors with the mRNA and the ribosome will require additional investigations with a wide range of biophysical/biochemical methods along with validation in both in vitro and in vivo assays. Here, we seek to elucidate the mechanisms of start codon recognition.

Studies on yeast factors have suggested a model for initiation factor-mediated conformational changes of the ribosome during the process of scanning and start codon recognition (Asano and Sachs, 2007; Hinnebusch, 2011).

(1) eIF1 and eIF1A stabilize the open complex with Met-tRNA$_i^{Met}$ loaded on the P-site. Here, the globular part of eIF1A occupies the A-site, whereas its C-terminal tail (CTT) extends into the P-site (Yu et al., 2009), and eIF1 binds on the 40S next to the Met-tRNA$_i^{Met}$ (Lomakin et al., 2003).
(2) eIF5 induces the hydrolysis of GTP bound to eIF2 as a GTPase-activating protein (GAP) upon or subsequent to mRNA binding to the PIC; however, the resulting GDP and P_i are thought to remain bound to eIF2 in the scanning PIC. P_i release is the rate-limiting step of the PIC in response to AUG selection.

(3) AUG recognition and tighter Met-tRNA$_i^{Met}$ binding trigger the transition to the closed state but require eIF1 dissociation and ejection of the CTT of eIF1A from the P-site.
(4) eIF1 dissociation enables gated release of P_i, which effectively ends the scanning mechanism. The dissociation of eIF1 with concomitant P_i release signifies the first irreversible step in translation initiation; hence, we asked whether eIF5-carboxyl terminal domain (CTD) serves a regulatory function in the ejection of eIF1 from PICs at start codons.

Capitalizing on different eIF1A dissociation kinetics between the open and closed states, individual mutations altering eIF1A or eIF5 (within its NTD [N-terminal domain]) were shown to manipulate the closed and open states of the AUG- or UUG-bound PIC (Maag et al., 2006). Any mutation that favors the closed PIC and allows initiation at a faulty UUG codon would promote translation initiation at the expense of lower fidelity, leading to a *su*ppressor of *i*nitiation (Sui$^-$) codon mutation phenotype, whereas a mutation that favors the open complex and blocks faulty initiation at a UUG codon would increase the fidelity of translation initiation, a suppressor of *Sui*$^-$ (Ssu$^-$) phenotype (Saini et al., 2010). Thus, Sui$^-$ mutations stabilize the closed state, whereas Ssu$^-$ mutations favor the open state of the PIC (summarized in Table S1 available online).

eIF5 is necessary for 60S ribosomal subunit joining, which is ultimately mediated by eIF5B and occurs only after start codon recognition and the cessation of scanning (Pestova et al., 2000). The best-characterized function of eIF5 is to serve, through its NTD (eIF5-NTD; residues 1–170), as the GAP for the eIF2-TC (Conte et al., 2006), which is critical for start codon recognition. The CTD of eIF5 plays a role in assembly of the PIC by stabilizing interactions among eIF1, eIF2, and eIF3 in the multifactor complex (MFC) (Asano et al., 2000; Sokabe et al., 2012). However, the interactions of eIF5-CTD with these other factors have not been fully characterized biophysically, and it was unknown whether the eIF5-CTD functions in scanning or AUG selection.

In order to elucidate critical initiation factor interactions required to promote start codon recognition and grasp a better understanding of the mechanism leading to the cessation of the scanning process, we employ an interdisciplinary approach to examine the dynamic function of the eIF5-CTD in conjunction with release of eIF1 from PICs, which signifies the first committed step in translation initiation. We identify overlapping surfaces on eIF5-CTD that bind to eIF1 and an N-terminal lysine-rich region of eIF2β. Using nuclear magnetic resonance (NMR) spectroscopy, small-angle X-ray scattering (SAXS), and isothermal titration calorimetry (ITC), we characterize these interactions and obtain a model for the eIF1:eIF5-CTD complex, which is congruent with the eIF1:40S structure. We unveil a common interface on eIF5-CTD that is important for start codon recognition and show that mutating these surface residues disrupts its binding to eIF1 and eIF2β. Surprisingly, our biochemical and genetic studies in yeast show that the CTD of eIF5 ends scanning by stabilizing the closed, scanning-arrested state of the PIC at start codons. We suggest that the NTD (GAP) and CTD (HEAT) of eIF5 serve a dual role in promoting the shift to the closed conformation in response to start codon selection. The eIF5-CTD interaction with eIF2β facilitates the subsequent release of eIF1 and ends the scanning process, while priming the PIC for eventual translation initiation.

RESULTS

Unveiling the Surface of eIF5-CTD that Binds to eIF1 and Characterization of This Interface

In order to map the surface of human eIF5-CTD that interacts with human eIF1, the chemical shift perturbation (CSP) assay was performed using uniformly ^{15}N-labeled human eIF5-CTD and unlabeled eIF1 (Figures S1A and S1B). We performed backbone resonance assignments of eIF5-CTD at 200 mM NaCl. Of the eIF5-CTD backbone resonance signals, 95% were assigned using the program IBIS, along with standard triple-resonance experiments (Hyberts and Wagner, 2003; Marintchev et al., 2007). eIF1-induced CSPs of eIF5-CTD backbone amide signals were mapped onto the surface of eIF5-CTD (Figures 1A–1D). We also used the NMR backbone resonance assignments of human eIF1 and the CSP assay (Fletcher et al., 1999) to map the human eIF5-CTD-binding surface on human eIF1 (Figures 1E, 1F, S1C, and S1D). The binding surface on eIF1's body includes a glycine (G101), along with both charged and hydrophobic residues (D53 and F63) (Figures 1E, 1F, and S1D).

We employed paramagnetic relaxation enhancement (PRE) measurements to generate distance restraints between eIF1 and eIF5-CTD. MTSL spin labels were attached to single-cysteine mutations of eIF1 (Figure S2A), and distance-dependent broadening of eIF5-CTD amide backbone resonances in complex with eIF1 was measured in ^{15}N-^{1}H TROSY HSQC spectra yielding distance restraints within a 20 Å radius (Figure S2B, green spectra). Addition of ascorbic acid reduces the MTSL nitroxide label restoring the intensity of the resonances previously broadened by the MTSL (Figure S2B, purple spectra). In Figures S2C–S2G, histograms show the intensity ratios of peaks in eIF5-CTD spectra in the presence (Figure S2B, green spectra) and absence of paramagnetic broadening (Figure S2B, purple spectra). This NMR PRE approach accomplishes two goals: (1) it provides distance restraints (<20 Å) between five positions on eIF1 and the multiple eIF5-CTD residues; and (2) it confirms that the surface on eIF5-CTD affected by eIF1 binding is at/near the binding interface. Distance restraints derived from the PRE experiments (Battiste and Wagner, 2000), along with the CSP data, were used in the HADDOCK software (Dominguez et al., 2003) to generate a model of the eIF1:eIF5-CTD complex (Figure 1G).

In order to validate the eIF1:eIF5-CTD model, we employed a site-directed mutagenesis approach. We examined whether changing the residues on the predicted eIF5-CTD surface would abrogate binding between eIF1 and eIF5. Based on our model (Figure 1G), we created a quadruple-mutant eIF5-CTD (eIF5-CTD-Quad; H305D/N306D/E347K/E348K), which remained folded upon incorporation of the four point mutations (Figure S1B, right panel). The eIF5-CTD-Quad mutant disrupts the binding of eIF5-CTD to wild-type (WT) eIF1, as shown by NMR CSP (Figures S1B and S1D, right panels). A panel of eIF5-CTD double mutants (H305A/N306A, H305D/N306D, E347A/E348A, and E347K/E348K) also maintained its folded state as assessed by

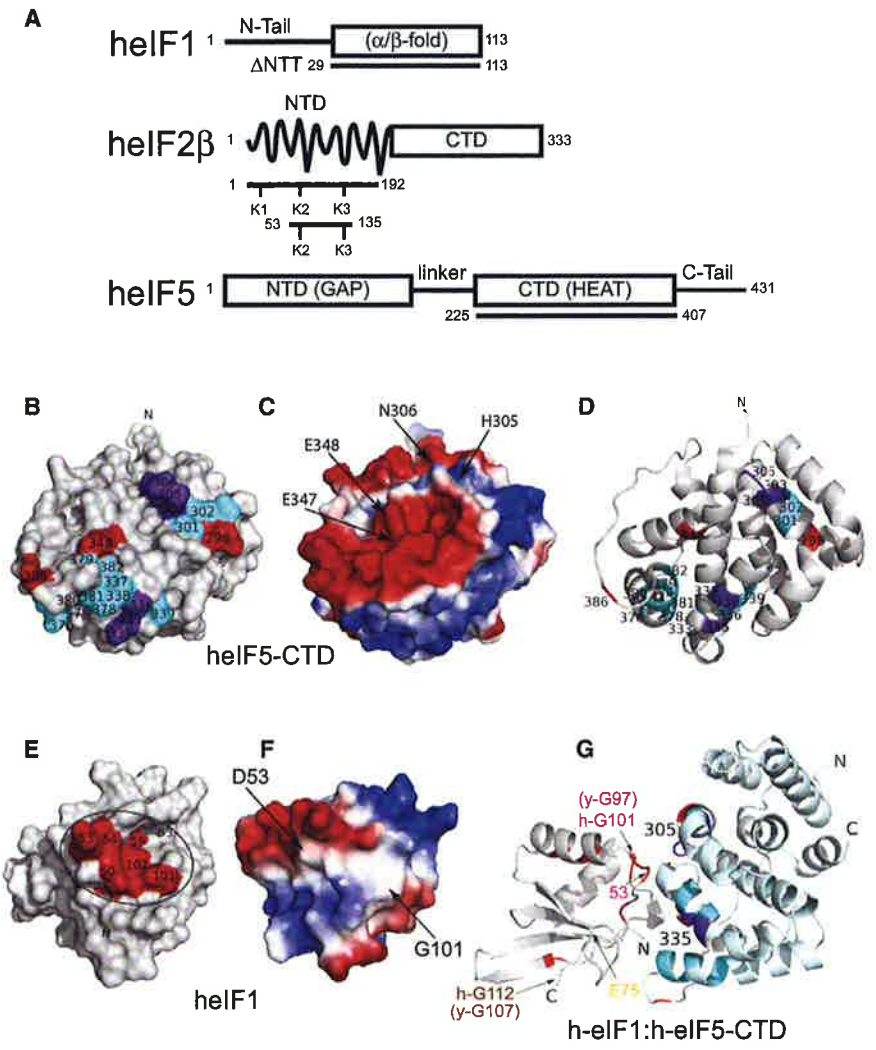


Figure 1. eIF5-CTD and the Globular Domain of eIF1 Form a Specific but Weak Binary Complex with a Well-Defined Interface

(A) Domain organization of eIF1, eIF2β, and eIF5 constructs used in this work. eIF5-CTD is a member of the HEAT domain family, *H*untingtin, eukaryotic elongation factor 3 (e*EF3*), protein phosphatase 2*A* and *TOR1* (target of rapamycin); α/β fold, eIF1 contains two α helices on one side of a five-stranded parallel and antiparallel β sheet.

(B) NMR mapping of the eIF1-binding surface on eIF5-CTD (2IU1). Contacts are only observed on one face of the domain. eIF5-CTD residues wherein eIF1 causes CSPs are painted red, residues that experience resonance broadening in PRE experiments are painted cyan, and residues affected in both experiments are painted purple. The figure seems to indicate two distinct binding patches based on monitoring backbone signals. As in helical proteins contacts are mainly made by side chains, which are more difficult to follow in chemical shift mapping experiments it is likely that the binding face is contiguous.

(C) eIF1-binding surface on eIF5-CTD showing the electrostatic potential of the surface of eIF5-CTD; similar orientation as (B) adjusted to show electrostatics.

(D) eIF1-binding surface on eIF5-CTD depicted as a ribbon diagram; same orientation as (B).

(E) eIF5-CTD-binding surface on eIF1 (2IF1) is circled. eIF1 residues wherein eIF5-CTD causes CSPs are painted red. The unstructured NTT of eIF1 is not shown for clarity.

(F) eIF5-CTD-binding surface on eIF1 showing the electrostatic potential of the surface of eIF1; similar orientation as (E) but adjusted to show electrostatics.

(G) Model for the eIF1:eIF5-CTD complex generated by the HADDOCK software from data summarized in (B), (E), and Figure S2. The same interface-coloring scheme is used as in (B), (D), (E), and (G). The location of one of the five single-cysteine mutants, eIF1-E75, is colored yellow. This residue was mutated to a single cysteine and used to attach a spin-label for PRE experiments. Human eIF1-G101 is colored red, and it is located within the interaction interface; this residue corresponds to yeast eIF1-G97. The location of human eIF1-G112 is colored brown, and it is not located within the interaction interface; this residue corresponds to yeast eIF1-G107.

See also Figures S1, S2, and S3.

NMR (Figure S3A). We examined whether double mutations on the surface of eIF5-CTD affect eIF1 binding; however, we were not able to completely abrogate binding between human eIF1 and eIF5-CTD using any of these double mutations as detected by NMR (Figure S3B). These mutagenesis results are congruent with the model in Figure 1G.

The Unstructured N-Terminal Lysine-Rich Tail of eIF2β Binds eIF5-CTD at an Epitope Overlapped with the eIF1-Binding Site

eIF2β-NTD contains three stretches of lysines, named K1, K2, and K3 boxes (Figure 1A). A previous study employing a panel of human eIF2β mutant constructs clearly shows that the NTD of eIF2β is responsible for binding to ^{32}P-labeled rat eIF5 used as a probe, wherein the K2 box of eIF2β was identified as the primary region for binding eIF5 (Das et al., 1997). Based on this finding, which we confirmed by NMR (Figures S4A–S4E), we produced a smaller eIF2β-NTD construct and named eIF2β-K2K3 (eIF2β residues 53–135; which contains the K2 and K3 boxes), and we mapped its binding surface on heIF5-CTD (Figures 1A, 2A, and 2B). Unlabeled eIF2β-K2K3 caused CSP and peak broadening of ^{15}N-labeled eIF5-CTD residues (Figures 2C and 2D). We noticed that the region of eIF5-CTD affected by eIF2β-K2K3 binding overlaps the region affected by eIF1 (compare Figures 1B and 2A). The eIF5-CTD-Quad mutant exhibited a drastically reduced ability to bind eIF2β-K2K3, as evidenced by significantly smaller CSPs and less broadening than observed with eIF5-CTD-WT (compare Figures 2D and 2E).

Using ITC, we measured an equilibrium dissociation constant (K_D) for the eIF5-CTD:eIF2β-NTD interaction of ~17 μM

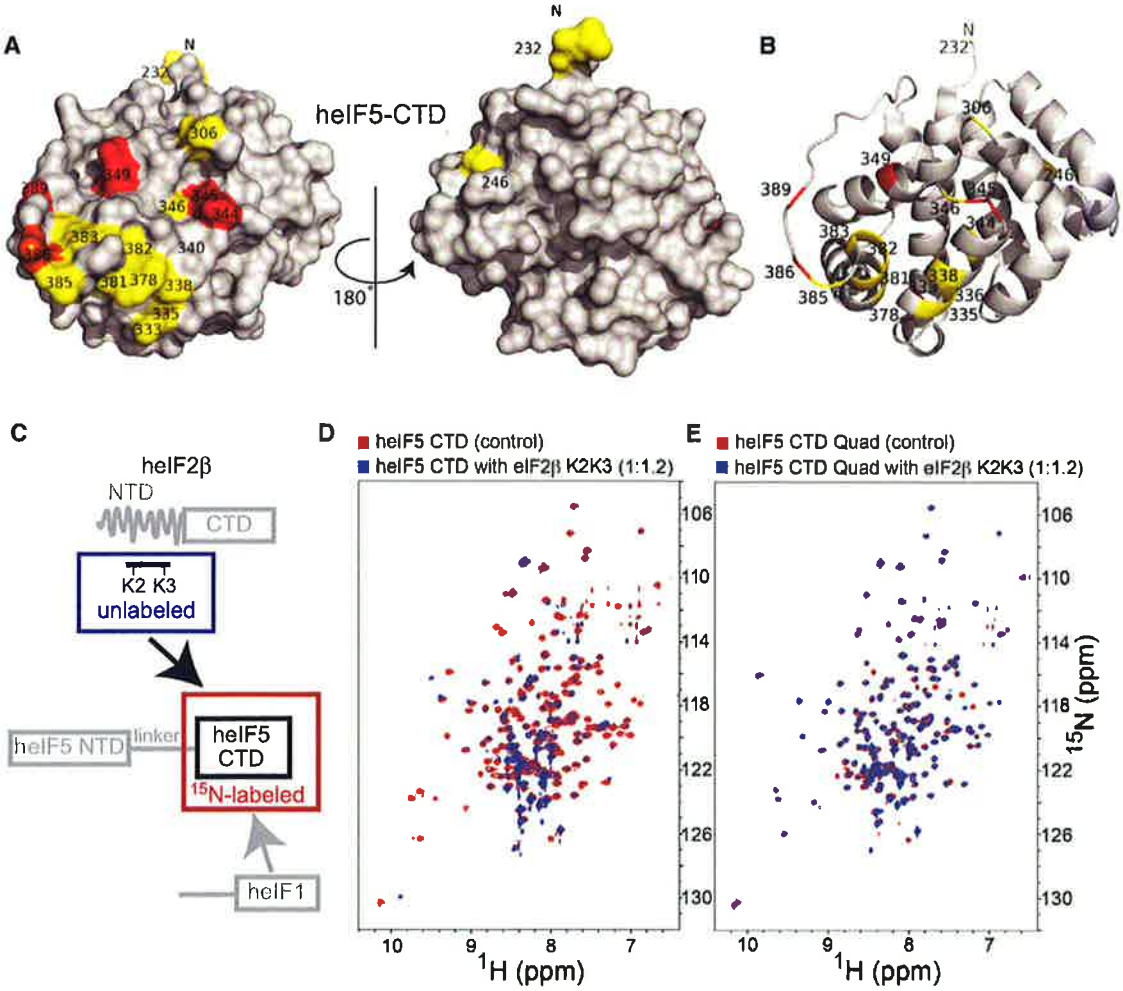


Figure 2. The eIF2β-K2K3 Segment Contacts eIF5-CTD on a Site Partially Overlapped with the eIF1-Binding Face

(A) eIF2β-K2K3-binding surface on eIF5-CTD. Two orientations are shown as surface representations; left orientation is the same as Figure 1B. Residues wherein eIF2β-K2K3 causes CSPs are painted red, those experiencing line broadening are painted yellow, and those seeing both effects are painted orange.

(B) eIF2β-K2K3-binding surface on eIF5-CTD depicted as ribbons; same orientation as left in (A).

(C) Schematic representation of the labeling scheme used in the proton-nitrogen correlation spectra: ^{15}N-labeled eIF5-CTD is measured alone (residues 225–407) (circumscribed in a red box) and in the presence of unlabeled eIF2β-K2K3 (residues 53–135) (circumscribed in a blue box).

(D) Overlay of ^{1}H-^{15}N HSQC spectra of 0.2 mM ^{15}N-labeled WT eIF5-CTD alone (red) and in the presence of 0.24 mM unlabeled eIF2β-K2K3 (blue).

(E) Overlay of ^{1}H-^{15}N HSQC spectra of 0.2 mM ^{15}N-labeled eIF5-CTD-Quad mutant domain alone (red) and in the presence of 0.24 mM unlabeled eIF2β-K2K3 (blue).

See also Figure S4.

(Figure S4F). The K_D for the eIF5-CTD interaction with the shorter eIF2β-K2K3 construct was measured as ~4 μM (Figure 3A). The similar eIF5-CTD-binding affinities with eIF2β-NTD and eIF2β-K2K3 confirm the findings in a previous eIF2β:eIF5 interaction study by Das et al. (1997) and validate our decision to utilize the eIF2β-K2K3 construct for mapping its interaction with eIF5-CTD. A noticeable baseline deviation after saturation (Figures 3A and 3B) does not contribute significantly to the qualitative comparisons among differing eIF5-CTD complexes. We found that the eIF5-CTD-DD mutant:eIF2β-K2K3 interaction exhibits a similar affinity, also in the low micromolar range, with a K_D of ~8 μM (Figure 3B). In contrast the binding affinity of eIF2β-K2K3 with either eIF5-CTD-KK or eIF5-CTD-Quad was abolished and could not be determined by ITC (data not shown), hence validating our NMR-chemical shift mapping data and substantiating the identification of a critical linkage between eIF5 and eIF2β.

SAXS Reconstitution Assay Shows that eIF5-CTD Does Not Simultaneously Bind eIF2β and eIF1

Because the eIF1:eIF5-CTD interaction could not be quantified by ITC measurements (data not shown), we employed SAXS (small-angle X-ray scattering). In this SAXS reconstitution assay, increasing amounts of eIF1 were titrated into a fixed concentration of eIF5-CTD, and the mixture at each point was subjected to X-ray analysis. The radius of gyration (R_g), being analogous to the

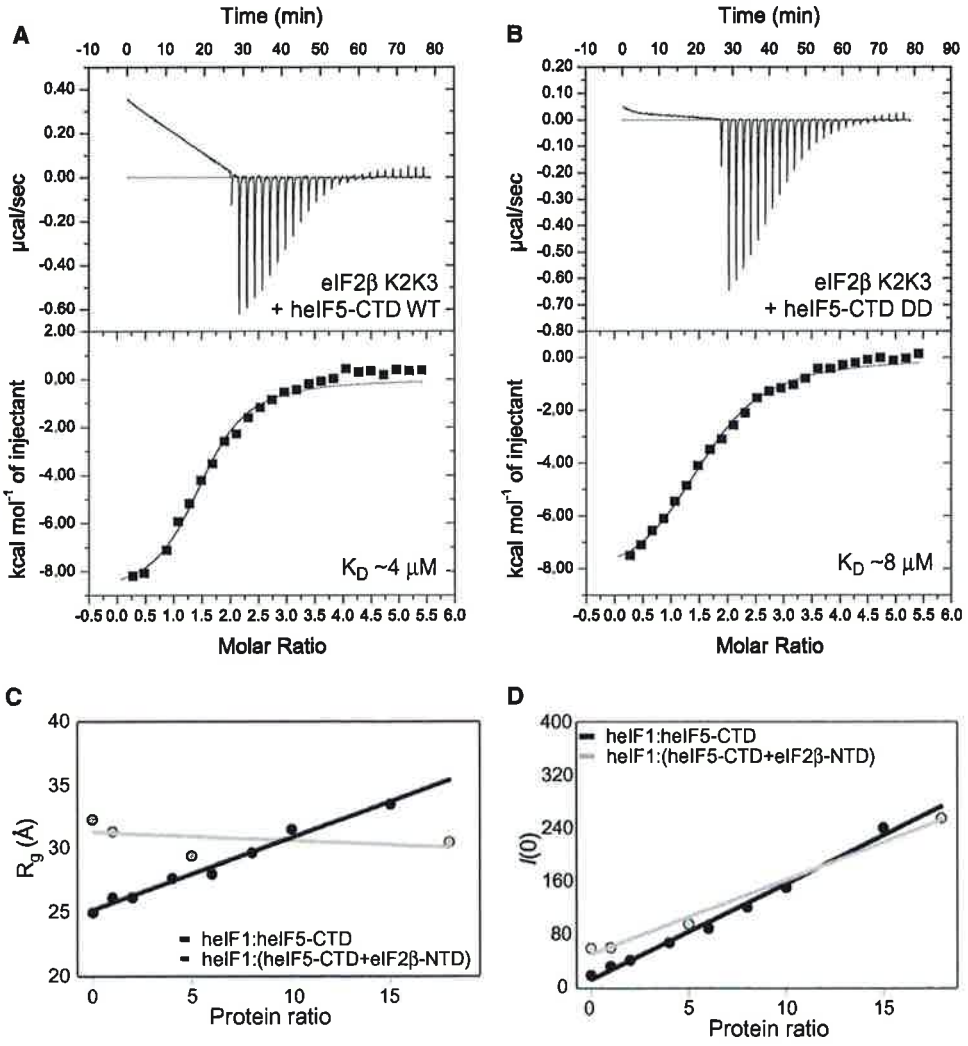


Figure 3. ITC and SAXS Narrow Down the eIF2β-Binding Site and Suggest Competitive Binding with eIF1

(A) ITC analysis of the eIF5-CTD:eIF2β-K2K3 interaction. A solution of eIF2β-K2K3 injected with eIF5-CTD WT.

(B) ITC analysis of the eIF5-CTD-DD mutant:eIF2β-K2K3 interaction. A solution of eIF2β-K2K3 injected with eIF5-CTD-DD. Interestingly, the KK and/or Quad(DDKK) mutations of eIF5-CTD abrogate binding to eIF2β.

(C and D) SAXS reconstitution assay used to monitor complex formations. Linear relation of the low-angle portion of the data corresponds to well-behaved proteins devoid of aggregation, even in samples with high concentration. (C) SAXS results plotting the R_g versus eIF1:eIF5-CTD (black) and eIF1:(eIF5-CTD+eIF2β-NTD) (gray) protein ratios. The data indicate that eIF1 binds eIF5-CTD but does not displace eIF2β-NTD or form a heterotrimeric complex. (D) Scattering intensities, *I*(0), shown as a function of concentration dependence; same protein samples as in (C). Plotting the scattering intensity (y axis) versus eIF1:(eIF5-CTD+eIF2β-NTD) molar ratios (x axis; gray lines) shows a concentration dependence. The same was done for eIF1:eIF5-CTD ratios (black lines). Data were collected for eIF5-CTD (90 µM) titrated with increasing amounts of eIF1. R_g serves as an indicator for the formation of higher-order protein complexes. SAXS results plotting the R_g (y axis) versus protein ratios (x axis). Black circles indicate eIF5-CTD (90 µM) titrated with increasing amounts of eIF1 (90, 180, 360, 540, 720, 900, 1,350, and 1,800 µM). Gray circles represent eIF5-CTD:eIF2β-NTD FPLC-purified complex (90 µM) titrated with increasing amounts of eIF1 (90, 450, and 1,620 µM).

See also Figure S5.

moment of inertia in mechanics, reflects the conformational/binding state of the proteins in solution, free or in complex with each other. Increasing the eIF1:eIF5-CTD ratio results in a steady increase in the R_g, consistent with complex formation between eIF1 and eIF5-CTD (Figure 3C, black spheres, and Figure S5A). However, there is no saturation of the R_g value, indicative of a weak-binding interaction between eIF1 and eIF5-CTD.

Because eIF1 and eIF2β-NTD bind to overlapping surfaces on eIF5-CTD, we used the SAXS reconstitution assay to monitor whether these three proteins bind simultaneously to form a higher-order complex in solution. Upon the titration of eIF1 to a preformed eIF5-CTD:eIF2β-NTD complex, these three proteins do not form a higher-order complex under these conditions, as evidenced by the lack of an increase in the R_g (Figure 3C, gray

spheres, and Figure S5B). The protein ratios in Figure 3D are the same as in Figure 3C, wherein the increase in the number of scatterers (amount of proteins) is related to the linear increase in the SAXS intensity: *I*(0) (Figure 3D, black and gray spheres).

The Interaction between eIF1 and eIF5-CTD Is Evolutionarily Conserved

Upon deletion of the flexible NTT of eIF1, the body of eIF1 still makes contact with eIF5-CTD near residues H305 and N306; other residues showing effects include D344, E348 (in the turn between helices 6 and 7), Y362 (in the turn between helices 7 and 8), and E386 (in the long loop between helices 8 and 9) (Figures S5C and S5D). Thus, the NTT of human eIF1 plays an ancillary role in binding to eIF5. As previously mentioned, the eIF1-binding surface on eIF5-CTD maps to an overlapping surface wherein eIF2β also binds (Figure S5E), and the mutated Quad residues directly impact this overlapping region (Figure S5F). We proceeded to examine heterologous interactions by NMR spectroscopy and found that human eIF5-CTD is able to bind to yeast eIF1, whereas the Quad mutation in human eIF5-CTD abolishes binding to yeast eIF1 (Figures S6A and S6B, middle and right panels). Importantly, the mapped contact surface on human eIF1 overlaps the previously mapped binding surface on yeast eIF1 for yeast eIF5-CTD (Figures S6C and S6D) (Reibarkh et al., 2008), involving three conserved human eIF1 residues: F13, A14, and G101 (corresponding to yeast F12, A13, and G97) (Figure S6C versus S6D). Figure S6E shows the sequences of human and yeast eIF1 with residues experiencing CSP upon interacting with human and yeast eIF5-CTD, respectively. Interestingly, the yeast *sui1-93–97* mutant exhibits a Sui$^-$ phenotype with a well-defined mechanism, involving accelerated eIF1 release from PICs (Cheung et al., 2007).

The Quad Mutation Impairs eIF5's Ability to Recruit eIF2-TC to PICs In Vitro

Previously, we identified eIF5's antagonistic interplay with eIF1 and eIF2-TC within PICs (Nanda et al., 2009). The binding of eIF1 and eIF1A to the 40S subunit promotes its open conformation, favorable for direct $tRNA_i^{Met}$ and mRNA binding to the decoding site (Passmore et al., 2007). In the presence of WT eIF1A, the yeast eIF1-G107K mutant protein diminished stable eIF2-TC loading on 40S subunits in vitro (without destabilizing mutant eIF1-G107K binding to the 40S) (Nanda et al., 2009). In this experimental situation (without mRNA), eIF2-TC loading on the 40S is contingent upon a shift to the closed conformation of the PIC, stabilizing $tRNA_i^{Met}$ bound to the P-site. eIF1 antagonizes this shift, and the eIF1-G107K mutant exhibits a stronger antagonistic ability due to its tighter binding to the 40S ribosomal subunit. Interestingly, this effect was overcome by the addition of yeast eIF5 (Nanda et al., 2009). Thus, eIF5 enhances eIF2-TC loading by promoting eIF1 release, which is consistent with our identification of conserved overlapping surfaces for eIF2β-K boxes and eIF1 in eIF5-CTD, along with our data establishing that human eIF5-CTD does not bind simultaneously to eIF1 and eIF2β-NTD in solution.

In this in vitro yeast system, we interpret the ability of the WT yeast eIF5 to enhance eIF2-TC binding to the PIC as a measure of its ability to promote eIF1 release. The Quad mutation introduced into the corresponding homologous eIF1/eIF2β-binding surface of yeast eIF5 reduces this eIF5-dependent rescue of eIF2-TC loading in reactions containing eIF1-G107K (Figure 4A). The yeast eIF5-Quad mutant exhibits significantly reduced ability to promote stable eIF2-TC recruitment relative to the WT factor, as evidenced by a 5-fold increase in the concentration of eIF5-Quad mutant required to achieve half-maximal eIF2-TC binding (Figure 4A, black curve, eIF5-Quad K_D = 0.55 ± 0.1 μM as compared to the red curve, eIF5-WT K_D = 0.10 ± 0.05 μM). Thus, the diminished ability of eIF5-Quad to enhance observable eIF2-TC loading in vitro is likely due to its impaired ability to promote the release of the eIF1-G107K mutant, which is bound tighter than eIF1-WT to 40S ribosomes.

The eIF5-Quad Mutant Destabilizes the Closed State of the PIC In Vitro

We proceeded to examine whether the Quad residues of eIF5 regulate the conformational change in the PIC response to AUG selection. Previous studies indicated that eIF1A dissociates more slowly from the PIC at AUG versus non-AUG codon (e.g., UUG) (Fekete et al., 2007; Maag et al., 2006). The in vitro eIF1A dissociation assay serves as a proxy for monitoring the opening/closing of the PIC upon start codon recognition. Stabilization of eIF1A binding upon start codon recognition is thought to indicate a closed state of the PIC, characterized by a reduced rate of eIF1A dissociation from the PIC in vitro, although the physiologically relevant dissociation of eIF1A takes place at the end of the initiation cycle, after subunit joining (Acker et al., 2006). 43S•mRNA(AUG) or 43S•mRNA(UUG) complexes were assembled with eIF1A that was labeled at its C terminus with fluorescein, in the presence of either eIF5-WT or eIF5-Quad mutant. These labeled complexes were then chased with excess unlabeled eIF1A, and dissociation of eIF1A (fluorescently labeled) was measured over time as a decrease in fluorescence anisotropy. Using eIF5-WT, eIF1A dissociates with biphasic kinetics, with rate constants for the fast and slow phases designated k_1 and k_2, respectively (Fekete et al., 2007; Maag et al., 2006). Previous studies indicated that the slow phase corresponds to eIF1A dissociation from PICs in the closed state, whereas the fast phase represents dissociation from complexes in the open state. K_{amp} is the ratio of the amplitudes of the slow to fast kinetic phases (an apparent equilibrium constant between the two states); hence, values of K_{amp} >1 indicate that the closed complex predominates in this system.

It was shown previously that the enhancement of eIF1A binding to PICs upon start codon recognition does not happen in the absence of eIF5 (Maag et al., 2006). Consistent with previous studies by Fekete et al. (2007) and Maag et al. (2006), the kinetics of eIF1A dissociation from AUG complexes in the presence of eIF5-WT is dominated by the slow phase (K_{amp} = 6.1 ± 1.5; Table 1, AUG, row 5) (Figure 4B, red line), whereas replacing AUG with UUG in the model mRNA reduces K_{amp} several fold (to 2.5 ± 0.5), indicating a stabilization of the open state relative to the closed state (Figure 4B, black line). This indicates that eIF5 promotes the shift to the closed state preferentially in response to AUG versus UUG. Dissociation of eIF1A from AUG complexes containing the eIF5-Quad mutant showed a 4-fold reduction in K_{amp} (Table 1, AUG, K_{amp} = 1.5 ± 0.5), such that

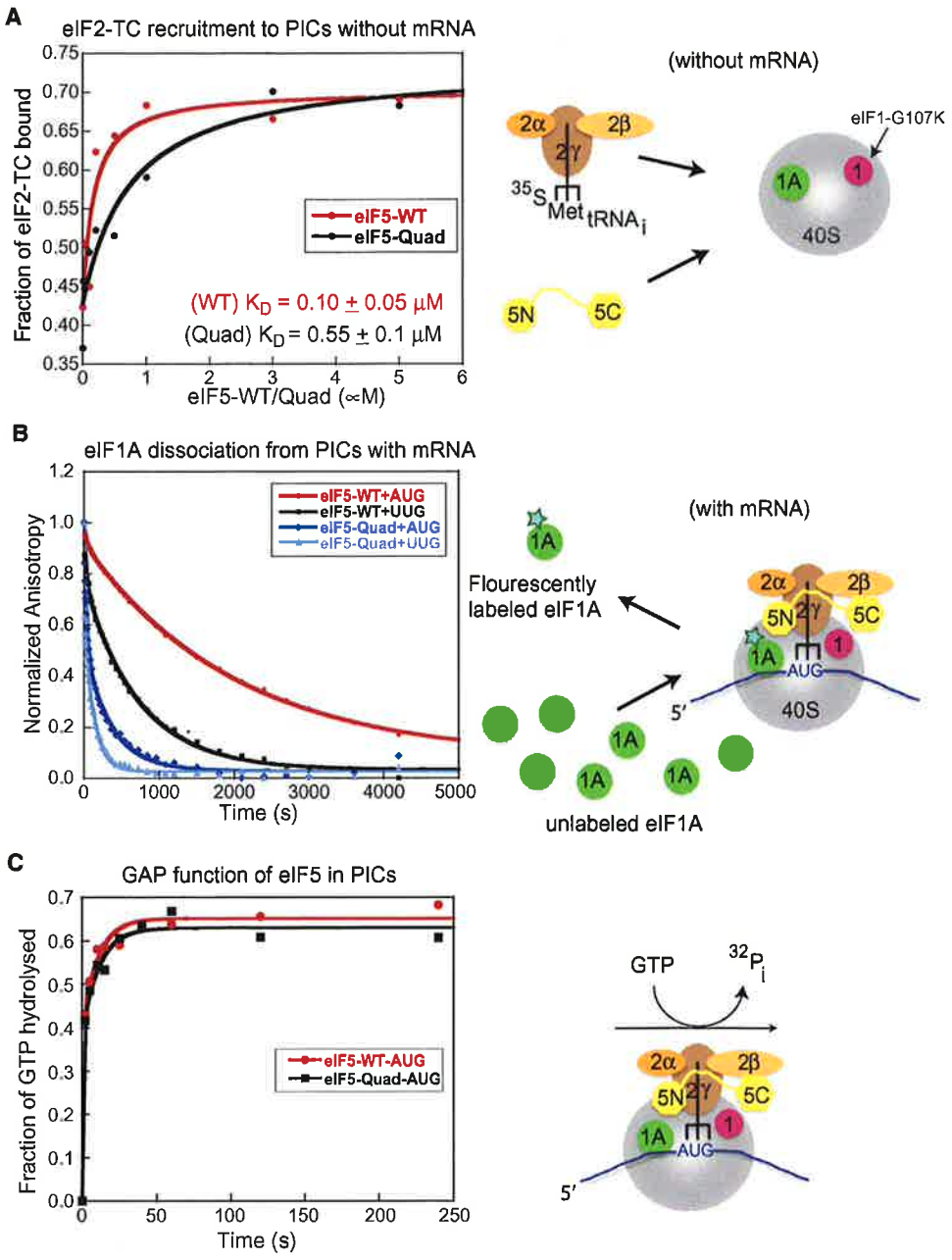


Figure 4. Biochemical Evidence that eIF5-CTD Functions to Promote eIF1 Release and the Closed Complex

These experiments used purified yeast eIFs and ribosomes, as summarized schematically on the right.

(A) eIF5-Quad mutant reduces eIF2-TC recruitment to 43S complex (−AUG) in vitro. eIF5-Quad mutant is unable to efficiently suppress the eIF2-TC recruitment defect conferred by the G107K eIF1 mutant (corresponds to human eIF1-G112) to the same extent obtained by WT eIF5. In this experiment using yeast eIF1-G107K mutant and WT eIF1A, eIF2-TC binding to 40S is severely compromised, but this effect is rescued by adding high amounts of WT eIF5 (red curve). A yeast eIF5-Quad mutant (V316D, T317D, S357K, E358K) is not able to rescue the eIF2-TC recruitment defect imposed by eIF1-G107K (black curve). In the case of WT eIF1, eIF5 is not needed for efficient eIF2-TC recruitment under these experimental conditions (data not shown).

(B) Effect of the eIF5-Quad mutant on eIF1A dissociation from 43S (AUG)/(UUG) complexes. The eIF5-Quad mutant destabilizes 43S (AUG)/(UUG) complexes, which leads to eIF1A dissociation from PICs. The kinetic constants of eIF1A dissociation from 43S AUG or UUG complexes from Figure 3B are shown in Table 1.

(C) Effect of the eIF5-Quad mutant on GTP hydrolysis from 43S.AUG complexes. The eIF5-Quad mutant has no significant effect on GTP hydrolysis (black squares) because the eIF5-Quad exhibits similar levels of hydrolysis when compared to WT eIF5 (red circles).

See also Figure S6.

Table 1. Kinetic Constants of eIF1A Dissociation from 43S AUG/UUG Complexes

eIF5 Allele	AUG	UUG
WT	k_1 = 9.0 ± 3.0; a_1 = 0.14 ± 0.06	k_1 = 7.0 ± 3.0; a_1 = 0.33 ± 0.1
	k_2 = 0.35 ± 0.05; a_2 = 0.86 ± 0.2	k_2 = 2.0 ± 0.5; a_2 = 0.67 ± 0.2
	K_{amp} = 6.1 ± 1.5 (a_2/a_1)	K_{amp} = 2.5 ± 0.5
Quad	k_1 = 30 ± 7.0; a_1 = 0.49 ± 0.05	k_1 = 60 ± 10; a_1 = 0.44 ± 0.1
	k_2 = 2.5 ± 0.5; a_2 = 0.51 ± 0.1	k_2 = 7.2 ± 1.0; a_2 = 0.56 ± 0.05
	K_{amp} = 1.5 ± 0.5	K_{amp} = 1.2 ± 0.2

All rates are 10^{-3} s^{-1}. k_1, fast-phase rate constant; a_1, fast-phase amplitude; k_2, slow-phase rate constant; a_2, slow-phase amplitude; K_{amp}, a_2/a_1 (closed/open). Fast-phase signifies open state. Slow-phase signifies closed state. K_{amp} is the ratio of closed to open state.

the rapid phase is increased in this reaction (Figure 4B, dark-blue line). For the corresponding UUG complexes, K_{amp} remains low, and k_1 and k_2 both are strongly increased by the Quad substitution (Table 1, Quad mutant; UUG K_{amp} = 1.2 ± 0.2) (Figure 4B, light-blue line), indicating that the Quad mutation should not confer the ability to initiate from the UUG codon. The Quad mutant's role as a GAP remains intact as evidenced by the ability to induce GTP hydrolysis similar to eIF5-WT (Figure 4C). Therefore, we conclude that the eIF5-Quad mutant destabilizes the closed state of the PIC relative to the open state, without affecting the GAP function. These findings led us to predict that the eIF5-Quad mutation will reduce the ability to access the closed conformation inappropriately at UUG codons in Sui$^-$ mutants, thereby conferring an Ssu$^-$ phenotype in yeast. This turns out to be the case, as described below.

eIF2β:eIF5-CTD Interaction Is Critical for Human eIF5 to Complement Yeast eIF5 Deletion In Vivo

We turned our attention to the in vivo function of the eIF5-CTD interaction network and assessed whether the mutations in the CTD exhibit a phenotype consistent with our biochemical and biophysical data. We employed an in vivo assay that exploits the ability of human eIF5 to provide the essential function of yeast eIF5 (Tif5). It was shown previously that mammalian eIF5 manifests the capability to functionally substitute for yeast eIF5 in a yeast strain deleted for the chromosomal gene encoding eIF5 (*tif5Δ*) (Maiti and Maitra, 1997). Here, we introduced into yeast the cloned coding sequences for either human eIF5-WT or the -DD, -KK, or -Quad mutants, tagged with the FLAG epitope and placed under the native *TIF5* promoter on a *LEU2* single-copy plasmid, and confirmed equal expression of eIF5-WT and mutant forms by western analysis (Figure 5A, bottom two gels). The plasmids were then tested for the ability to replace a single-copy *URA3 TIF5* plasmid in a *tif5Δ* strain using the drug 5-fluoro-orotic acid (5FOA) to select against *URA3* (plasmid shuffling). The plasmid encoding FLAG-tagged human eIF5 harboring the DD substitutions (heIF5-DD-FLAG) complemented *tif5Δ* similar to those encoding human eIF5-WT untagged or FLAG tagged (Figure 5A, rows 1–3). By contrast the constructs encoding the KK or Quad variants of FLAG-tagged heIF5 did not complement *tif5Δ* (Figure 5A, rows 4 and 5). Thus, the KK or Quad substitutions, but not the DD substitutions, abolish the ability of human eIF5 to provide the essential function of eIF5 in yeast cells. Because the Quad mutation is required to disrupt binding of human eIF5-CTD to eIF1 (Figures S1B, S1D, S3A, and S3B), whereas both the KK and Quad mutations abolish eIF5-CTD binding to eIF2β-NTD; hence, we conclude that the essential partner of human eIF5-CTD in yeast cells is likely to be eIF2β-NTD.

The KK and Quad Mutations in Yeast eIF5-CTD Reduce Aberrant Initiation from a UUG Start Codon In Vivo

We proceeded to evaluate the effects of introducing the DD, KK, or Quad mutations into yeast *TIF5* gene on the accuracy of start codon selection in vivo. To this end, we constructed these *TIF5* mutants in the *tif5Δ* strain harboring the *his4-306* allele of the *HIS4* gene in which the AUG start codon is altered, and translation begins on a downstream UUG codon. The strain containing WT *TIF5-FL* cannot grow on medium devoid of histidine (Figure 5B, row 1, −His medium) because the start codon mutation in *his4-306* abolishes expression of this histidine biosynthetic enzyme. A strain containing *TIF5-G58S* altering the eIF5-NTD can grow on −His medium (His$^+$ phenotype) because an in-frame UUG triplet at the third codon of *his4-306* can be used as the initiation codon in cells harboring this eIF5 Sui$^-$ mutation (Singh et al., 2005) (Figure 5B, row 2). The strain harboring *TIF5-Quad* did not display a slow-growth (Slg$^-$) phenotype on +His medium (Figure 5B, row 3, +His), or a His$^+$/Sui$^-$ phenotype on −His medium (Figure 5B, row 3, −His). The absence of a Sui$^-$ phenotype (Sui$^+$) for the *TIF5-Quad* mutant is consistent with our in vitro analysis of eIF1A dissociation kinetics, which revealed a predominantly open state of 43S.UUG complexes harboring the yeast eIF5-Quad mutant (Figure 4B; Table 1).

As noted above, the analysis of eIF1A dissociation kinetics led us to predict that the *TIF5-Quad* mutant would display an Ssu$^-$ phenotype, suppressing the relaxed stringency of start codon recognition conferred by a Sui$^-$ mutation. To test this prediction, we asked whether the *TIF5-Quad* mutation can suppress the dominant His$^+$/Sui$^-$ phenotype conferred by the *SUI3-2* mutation in eIF2β (S264Y substitution), which appears to result from elevated eIF5-independent GTPase activity by the eIF2-TC and a weakened interaction between Met-tRNA$_i^{Met}$ and eIF2-GTP (Huang et al., 1997). As expected, introducing plasmid-borne *SUI3-2*, but not WT *SUI3*, into the *his4-306* strain harboring WT *TIF5* confers growth on the −His medium (Figure 5C, rows 1 and 2, −His). Furthermore, the dominant His$^+$/Sui$^-$ phenotype of *SUI3-2* is partially suppressed in the strain harboring a known Ssu$^-$ allele of *TIF5*, *tif5-G62S* (Asano et al., 2001), because only weak growth occurred on −His medium even after 5 days of incubation (Figure 5C, compare row 9 to row 1, −His). In agreement with our prediction, the *TIF5-Quad* allele confers an Ssu$^-$ phenotype nearly as strong as that of *tif5-G62S* (Figure 5C, compare rows 1, 7, and 9, −His). Importantly, the *tif5-KK* allele also confers a marked Ssu$^-$ phenotype, only slightly less than that of *tif5-Quad*, whereas the *tif5-DD* allele displays little or no Ssu$^-$ phenotype (Figure 5C, rows 3, 5, and 7; −His). The fact that the *TIF5-KK* mutation confers an obvious Ssu$^-$ phenotype

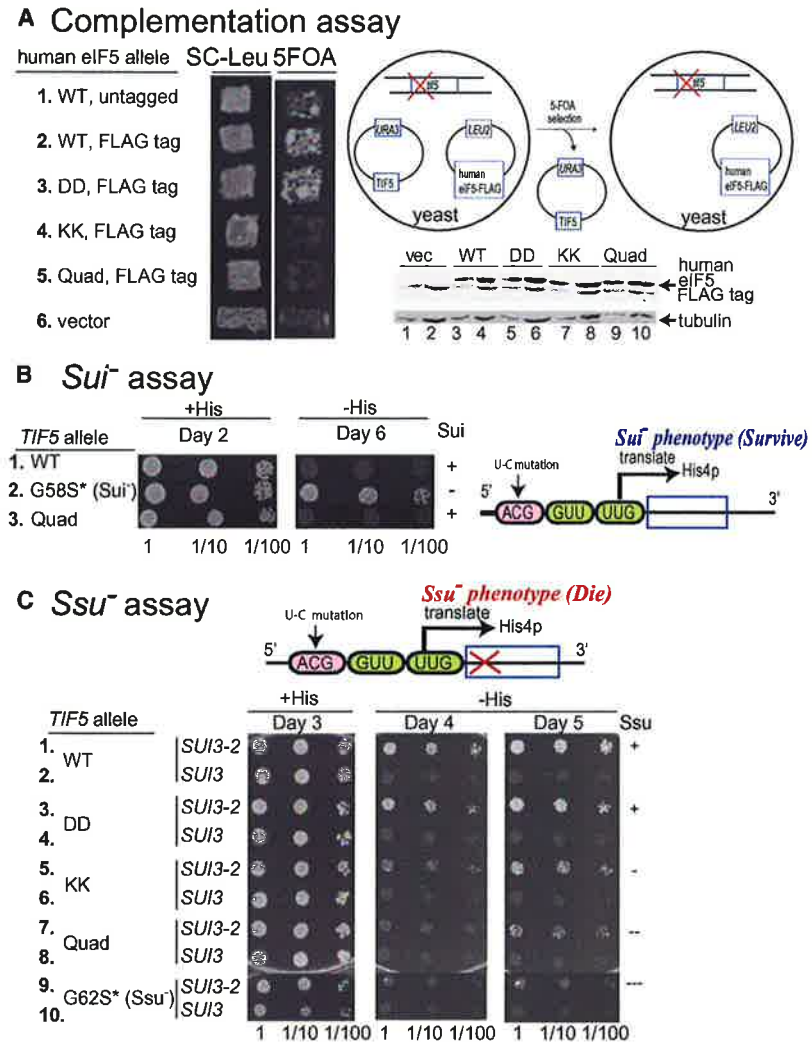


Figure 5. Genetic Evidence that the KK and Quad Substitutions Disrupt a Critical Linkage with eIF2β and Impair the Function of eIF5-CTD in Promoting the Closed Complex

(A) Complementation of *tif5Δ* by single-copy human eIF5 plasmids in yeast: 1., YCpL-heIF5; 2., YCpL-heIF5-FLAG; 3., YCpL-heIF5-DD-FLAG; 4., YCpL-heIF5-KK-FLAG; 5., YCpL-heIF5-DDKK-FLAG; and 6., YCplac111. Left view shows that patches of transformants of KAY24 (*tif5Δ* p [TIF5 URA3]) carrying the indicated plasmids were replica plated onto SC-leu agar plates and SC agar plates containing 5-FOA and uracil, which was used to evict the residing TIF5 URA3 plasmid. Cells were allowed to grow on these plates for a few days. The schematic of FOA-negative selection is shown on the right. A total of 20 and 40 μg (odd and even numbered lanes, respectively) of cell extracts from the KAY24 transformants carrying YCplac111 (vec) or the indicated YCpL-heIF5-FLAG derivatives (Table S4) was subjected for immunoblotting with antibodies indicated to the right. Upper panel shows westerns using anti-Flag antibody to detect the relative expression levels of FLAG-tagged human eIF5 proteins (top bands; bottom bands are cross-reactivity, also showing equal loading). Lower panel illustrates western blotting results showing comparable expression levels of a housekeeping protein, tubulin.

(B) Test of Sui⁻ phenotypes. *TIF5* alleles correspond to eIF5 alleles in yeast. A lack of Sui⁻ phenotype is represented by "+." Cultures (A_{600} = 0.15) of KAY976 (*his4-306*) derivatives (Table S5) with indicated *TIF5* alleles and their dilutions indicated on bottom are spotted onto SD+Trp+Ura medium with (+His) or without (−His) and incubated for indicated number of days.

(C) Test of Ssu⁻ phenotypes against *SUI3-2* (eIF2β-S264Y). A lack of Ssu⁻ phenotype is represented by "+." KAY976 (*his4-306*) derivatives (Table S5) with indicated *TIF5* alleles carrying YCpU-SUI3-2 (*SUI3-2*) or YCpU-SUI3 (*SUI3*) control plasmid (Watanabe et al., 2010) were grown and spotted onto SC-Ura (+His) or SC-Ura-His (−His) as in (B) and incubated for indicated number of days. The asterisk (*) following G58S or G62S represents the numbering for yeast residues located in the NTD of *TIF5*.

See also Tables S1, S2, and S3.

suggests that the weakened binding of eIF5-CTD to eIF2β produced by the KK substitution is responsible for this phenotype. These findings provide strong support for the conclusion that disrupting the interaction of eIF5-CTD with eIF2β destabilizes the closed conformation of the PIC, increasing the stringency of start codon recognition and thus decreasing initiation at near-cognate UUG codons in vivo.

In Vivo Evidence that the KK and Quad Substitutions in eIF5-CTD Confer Ssu⁻ Phenotypes by Decreasing eIF1 Release from the PIC

Our biochemical analysis of the eIF5-Quad mutant showed a reduction in eIF2-TC recruitment to PICs, which is indicative of a defect in promoting eIF1 dissociation upon AUG recognition. The previously described G107R and G107K eIF1 mutants exhibit the same defect in eIF2-TC recruitment. The G107R/K substitutions increase the UUG:AUG ratio (Sui⁻ phenotype) primarily by decreasing initiation at AUG rather than elevating UUG initiation (Nanda et al., 2009). By contrast, other eIF1 mutations, notably the 93–97 substitution in helix α2 (Cheung et al., 2007), appear to increase the UUG:AUG ratio by provoking more rapid dissociation at UUG codons owing to weaker eIF1 binding to the PIC. Accelerated eIF1 release likely also explains the Sui⁻ phenotype of *sui1-K60E* (Martin-Marcos et al., 2011) because Lys-60 contacts 18S rRNA in the eIF1:40S crystal structure (Rabl et al., 2011), and overexpression of *sui1-K60E* suppresses its Sui⁻ phenotype in the manner observed previously for *sui1-93–97* (data not shown). Because the Quad substitution impairs eIF1 dissociation, we reasoned that it should not suppress the elevated UUG:AUG ratio conferred by the G107R/K substitutions because they also impede eIF1 dissociation but that it should suppress the Sui⁻ phenotypes of the 93–97 and K60E substitutions by mitigating the accelerated eIF1 dissociation they engender.

We tested these predictions by examining a set of *sui1Δ his4-301* yeast strains (with a different start codon mutation at

Cell PRESS
Open ACCESS

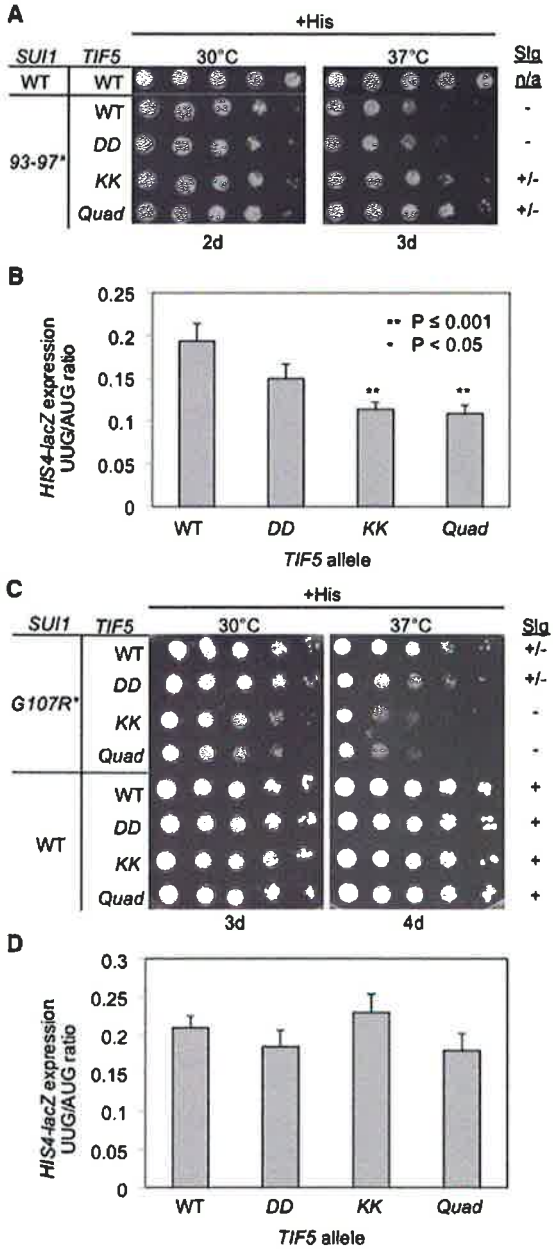


Figure 6. Genetic Evidence that the KK and Quad Substitutions Impair eIF5-CTD Function in Releasing eIF1 from the Closed Complex

SUI1 and *TIF5* alleles correspond to yeast *eIF1* and *eIF5* alleles, respectively.

(A and B) The *TIF5 KK* and *Quad* mutations suppress the Slg$^-$ and Sui$^-$ phenotypes of *sui1-93–97*. A lack of Sui$^-$ or Slg$^-$ phenotype is represented by "+." Derivatives of *sui1Δ his4-301* P_{GAL}*-TIF5* strain PMY01 harboring *sui1-93–97* and the indicated plasmid-borne *TIF5* alleles, plus control strain PMY106 containing WT *SUI1* and WT *TIF5*, were spotted in 10-fold serial dilutions on synthetic dextrose complete (SC) medium supplemented with 0.3 mM histidine (+His) or lacking His (−His) and incubated for the indicated number of days. (B) Strains from (A) also harboring *HIS4-lacZ* reporter plasmids with an AUG (p367) or UUG (p391) start codon were cultured in synthetic dextrose minimal medium (SD) supplemented with His at 30°C to A_{600} of ~1.0, and β-galactosidase activities (nanomoles of o-nitrophenyl β-D-galactopyranoside cleaved per min per mg) were measured in whole-cell extracts (WCEs). The ratio of expression of the UUG versus AUG reporter was calculated for replicate experiments, and the mean and SEM (error bar) were plotted. The UUG/AUG ratios were determined to be significantly smaller in the *tif5-KK* and *-Quad* mutants versus WT *TIF5* strain by the Student's t test.

(C and D) The *KK* and *Quad* mutations exacerbate the Slg$^-$ phenotype and do not suppress the elevated UUG:AUG ratio, conferred by *sui1-G107R*. Transformants of the indicated genotype were examined for growth on SC +His medium in (C), just as in (A), and for the UUG:AUG initiation ratio for *HIS4-lacZ* in (D), just as in (B). The UUG/AUG ratios were not significantly different among the WT and mutant *TIF5* strains. The asterisk (*) following *SUI1* mutants (*93–97* and *G107R*) represents the numbering for residues located in yeast eIF1. The corresponding human eIF1 residues are the following: *V98-G101* and *G112*.

See also Figure S7.

HIS4) containing plasmid-borne *TIF5* and *SUI1* alleles encoding the eIF5 and eIF1 mutants of interest and the chromosomal *TIF5* allele placed under the *GAL1* promoter to enable its repression on glucose medium. As expected, the WT *TIF5* strain expressing eIF1-93–97 has a Slg$^-$ phenotype at 37°C on +His (glucose) medium but can grow on media containing no histidine, indicating the Sui$^-$ phenotype (Cheung et al., 2007). Interestingly, the *TIF5-KK* and *TIF5-Quad* alleles improve the growth of *sui1-93–97* cells on +His medium but diminish their growth on −His medium, whereas the *TIF5-DD* allele only reduces growth on −His (Figure 6A). These results suggest that the *KK* and *Quad* mutations suppress the Sui$^-$ phenotype of *sui1-93–97* cells more effectively than does *TIF5-DD*. Supporting this, the *KK* and *Quad* mutations exceed *DD* in decreasing the UUG:AUG initiation ratio in *sui1-93–97* cells measured using *HIS4-lacZ* fusions with an AUG or UUG start codon. Thus, the *KK* and *Quad* mutations reduce the UUG:AUG ratio by ≈40%, similar in magnitude to reductions observed for Ssu$^-$ substitutions in eIF1A (Saini et al., 2010), whereas the *DD* mutation produces only a ≈20% reduction in the ratio. Importantly, the *KK* and *Quad* mutations also exceed the *DD* mutation in reducing the His$^+$/Sui$^-$ phenotype and elevated UUG:AUG ratio in *sui1-K60E* cells (Figures S7A and S7B).

The G107R substitution in eIF1 increases the UUG:AUG ratio to the same extent as does the 93–97 substitution (compare Figures 6B and 6D). Importantly, however, none of the CTD mutations of eIF5 produces a significant reduction in this ratio (Figure 6D). In addition the *KK* and *Quad* mutations exacerbate the Slg$^-$ phenotype of *sui1-G107R* cells (Figure 6C), opposite the effect displayed in *sui1-93–97* cells, and similar findings were observed for the mechanistically

related G107K mutation (Figures S7C and S7D). These findings support our prediction that the *KK* and *Quad* substitutions would not suppress the elevated UUG:AUG initiation ratio and would intensify the defect in eIF1 release on AUG recognition, conferred by eIF1-G107R/K. These results provide in vivo evidence that the *KK* and *Quad* substitutions impair the eIF5-CTD function in releasing eIF1 from the closed complex upon start codon recognition. Our molecular interpretations of the Ssu^- and Sig^- phenotypes of the various mutants harboring the eIF5-CTD mutations are shown in Tables S2 and S3, respectively.

DISCUSSION

eIF1 binds in the vicinity of the P-site on the 40S subunit and contacts Met-$tRNA_i^{Met}$ (Lomakin et al., 2003; Rabl et al., 2011). Therefore, we sought to use the eIF1:eIF5-CTD model to infer the position of eIF5-CTD in the PIC, by superimposing the eIF1:eIF5-CTD complex onto the eIF1:40S structure (Rabl et al., 2011) by aligning the common component: eIF1 (Figures 7A and 7B). The resulting model places eIF5-CTD in the vicinity of the Met-$tRNA_i^{Met}$, consistent with the eIF5-NTD binding to eIF2γ near the 3′ end of the tRNA (Alone and Dever, 2006; Conte et al., 2006).

The conclusion that the CTD of eIF5 promotes eIF1 dissociation is supported by our findings that the Quad substitution impairs the ability of eIF5, when added in excess, to reverse the antagonistic effect of eIF1-G107K on eIF2-TC binding in the reconstituted system. The mechanism wherein overlapping surfaces on the CTD of eIF5 are utilized to promote eIF1 dissociation was not known. We provide biochemical evidence that the CTD of eIF5 plays a previously unknown role in the promotion of eIF1 dissociation upon AUG recognition via its dynamic interaction with eIF1 and eIF2β, which is based on the following two findings: (1) the Quad substitution impairs eIF5's ability to promote stable recruitment of eIF2-TC to PICs containing eIF1-G107K (which binds with greater affinity to 40S ribosomes) without mRNA in vitro; and (2) the Quad substitution destabilizes closed PIC conformation with mRNAs upon start codon recognition. It is important to note that the role of eIF5-CTD in promoting eIF1 release is independent of the eIF5 GAP function catalyzed by the NTD.

A key question is whether the newly identified function of the CTD of eIF5 in promoting eIF1 release involves both of its interactions with eIF1 and eIF2β. We show here that the KK substitution on the CTD of eIF5 is sufficient to drastically reduce its binding to eIF2β, but not to eIF1. Thus, it can be deduced that disrupting the interaction of eIF5-CTD with eIF2β is likely to be responsible for destabilizing the closed PIC conformation and conferring the Ssu^- phenotypes displayed by the *eIF5-KK* and *eIF5-Quad* mutations. In fact it was shown previously that elimination of K boxes 1 and 2 from eIF2β also confers an Ssu^- phenotype (Laurino et al., 1999). Hence, it seems likely that the Ssu^- phenotypes produced by these newly identified substitutions in the CTD of eIF5 (along with the previously identified eIF2β-NTD Ssu^- phenotype) directly impact the linkage of eIF2β with eIF5 within the PIC.

It is plausible that the eIF1:eIF5-CTD complex is stabilized by the mutual interaction of these factors with eIF3c, which was demonstrated in vitro for the cognate yeast proteins (Asano et al., 2000). We propose that the eIF1:eIF5-CTD interaction is redundant with other interactions that stabilize eIF1 binding to the 40S leading to an open PIC conformation; hence, eliminating this interaction is probably not sufficient to accelerate eIF1 dissociation at UUG codons in vivo. Our conclusion that eIF5-CTD stabilizes the closed conformation of PICs upon start codon recognition via its interaction with eIF1 and eIF2β was based on the cumulative breadth of data ranging from NMR spectroscopy, ITC, SAXS, and in vitro yeast-reconstituted system. The in vitro assays employed in this study, which include eIF1, eIF1A, eIF2-TC, eIF5, and 40S ribosomes (with and without mRNA), utilize the basic components for start codon recognition and have been consistently implemented as useful tools to decipher biochemical events surrounding start codon recognition (Maag et al., 2006; Nanda et al., 2009). Based on our in vitro assays, we propose that eIF5 interactions, particularly with its CTD, are sufficient to promote eIF1 release by stabilizing the closed complex via its interaction with eIF2β. Our hypothesis was further examined using in vivo experiments, wherein we dissect the cause of Ssu^- phenotypes produced by mutations in the CTD of eIF5. We show that the *KK* and *Quad* mutants exhibit strong Ssu^- phenotypes, which further substantiates the role of eIF5-CTD in facilitating the stabilization of the closed state of PICs upon start codon recognition via its dynamic interplay with eIF1 and eIF2β.

The common interface on eIF5-CTD that contacts eIF2β also interacts with eIF1, in a manner disrupted by the Quad substitution. eIF1 and eIF2β do not form a higher-order complex with eIF5-CTD in our SAXS reconstitution assay; hence, an intriguing model would be that eIF5-CTD first interacts with eIF1 in the open, scanning conformation of the PIC and then switches partners to interact with eIF2β in the closed conformation upon AUG recognition (Figure 7C). In this view the interaction network (including eIF5-CTD) positions eIF1 in the decoding site and mediates scanning. We propose the following possible model of events surrounding start codon recognition. eIF5-CTD binds to eIF2β, which in turn favors the release of eIF1 from the PIC and essentially puts an end to the ribosomal scanning mechanism. This would also impede eIF1's reassociation with the 40S and drive the conformational rearrangement of the PIC to the closed state.

The hypothesis that the disruption of the eIF1:eIF5-CTD interaction by eIF2β triggers the critical shift from the open state (scanning competent) to the closed state (scanning incompetent) led us to conclude that the CTD of eIF5 stabilizes the closed state of the PIC in which eIF1 is no longer bound (Figure 7C). This model is fully consistent with our biophysical, biochemical, and genetic data wherein we strategically disrupt the eIF5-CTD:eIF2β interaction that destabilizes the closed state of the complex and prevents eIF5-CTD from promoting eIF1 release. We do note that in vivo other initiation factors, e.g., possibly eIF1A, eIF3, or eIF4G, could also play contributing roles in the observed Ssu^- phenotype caused by mutations in the CTD of eIF5. This study opens an avenue for further investigation to continue deciphering the effect of eIF5-CTD's interaction with its cognate partners and its crucial role in mediating the closure of PICs upon start codon recognition.

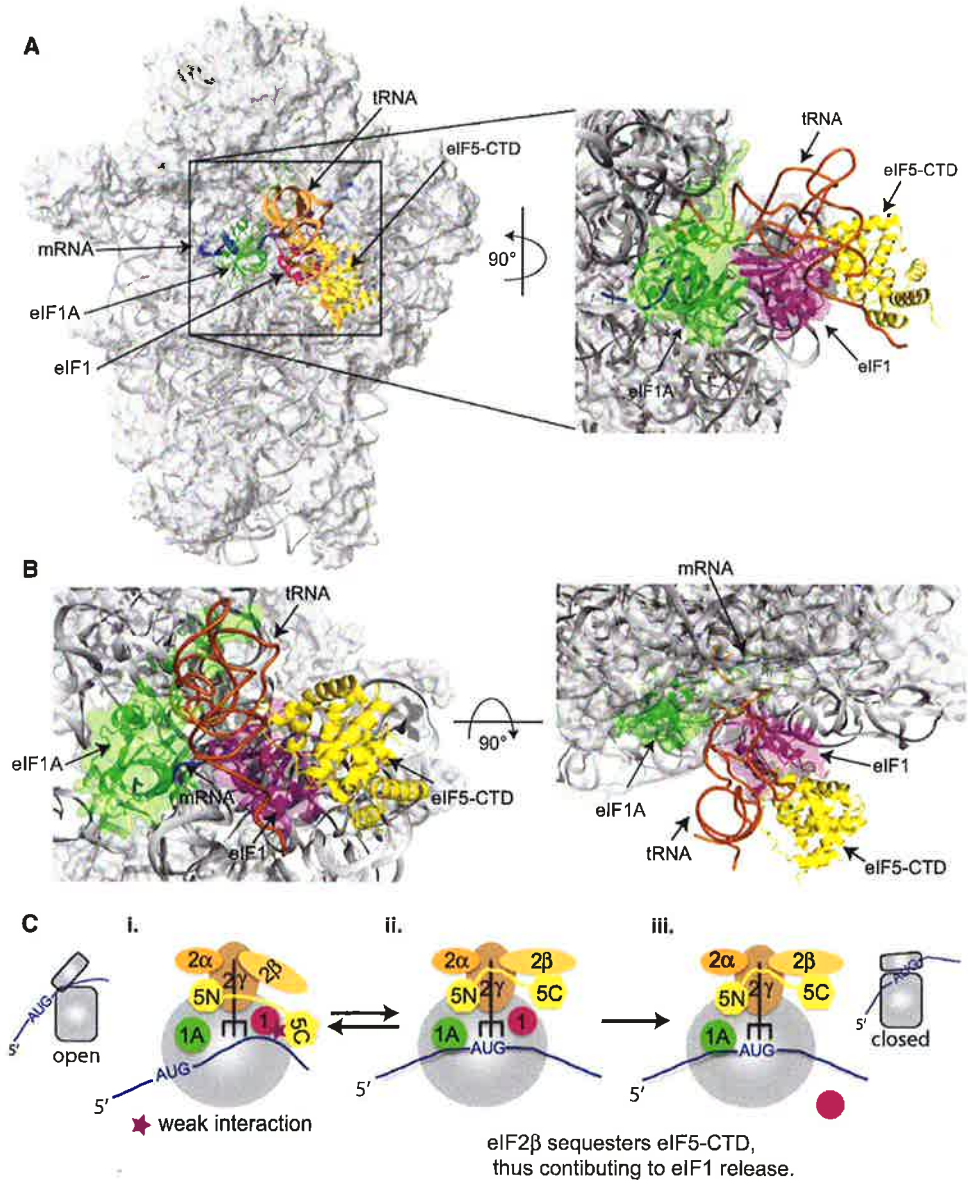


Figure 7. Modeling the eIF1:eIF5-CTD Complex in the Context of the eIF1:40S Structure and Schematic Showing the role of eIF5-CTD in Promoting Start Codon Recognition

(A) The eIF1:eIF5-CTD complex derived from the HADDOCK software was superimposed on the position of human eIF1 on the small ribosomal subunit (Lomakin et al., 2003; Rabl et al., 2011), shown as a magenta ribbon. The modeled position of the human eIF5-CTD (yellow) was obtained by aligning eIF1 from the eIF1:eIF5-CTD complex (Figure 1G) to ribosome-bound eIF1. eIF1A (green), mRNA (blue), and P-site tRNA (orange) are also shown for reference. The P-site tRNA is shown in a P/P orientation. Note that the actual orientation of the $^{Met}tRNA_i$ on the 40S subunit cannot be exactly in a P/P orientation because it would clash with eIF1 (Rabl et al., 2011) and is likely to change over the course of translation initiation. In the right panel the modeled eIF1A:eIF1/eIF5-CTD:40S ribosomal complex was rotated 90° along the y axis and magnified.

(B) In the left panel the eIF1A:eIF1/eIF5-CTD:40S ribosomal complex is in the same orientation as the left panel of (A) but magnified. In the right panel the modeled eIF1A:eIF1/eIF5-CTD:40S ribosomal complex was rotated 90° along the x axis from the left panel.

(C) Schematic diagram of eIF5-CTD gating the release of eIF1 after start codon selection, followed by subsequent stabilization of the PIC by eIF5. (i.) The open 43S conformation allows for mRNA recruitment and relatively unstable eIF2-TC binding. During the assembly stage of the PIC, eIF5-CTD (HEAT) interacts with eIF1. (ii.) The 43S PIC scans the mRNA in an open conformation until start codon recognition. During this scanning stage the eIF5-NTD induces eIF2γ to cleave GTP. Start codon recognition sets the stage for large conformational rearrangements on the subunit interface ensuring the closed state of the PIC. eIF2β exhibits a stronger affinity for the overlapping binding surface of the eIF5-CTD than eIF1; hence, the disruption of the eIF1:eIF5-CTD interaction by eIF2β allows for an indirect mechanism to dislodge eIF1 from the 43S PIC. Upon release of eIF1 the free phosphate is subsequently released. (iii.) eIF5-CTD stabilizes the closed ribosomal conformation of PICs upon start codon selection.

EXPERIMENTAL PROCEDURES

Biophysical Experiments and Analyses

NMR experiments were performed as described previously by Marintchev et al. (2007) and are written in more detail in the Extended Experimental Procedures. We recorded a set of triple resonance experiments needed for backbone assignments of human eIF5-CTD. Assignments were completed for 95% of the residues in the human eIF5-CTD, using standard techniques and IBIS (Hyberts and Wagner, 2003). Chemical shift mapping was done as previously described by Marintchev et al. (2007). A paramagnetic spin-label strategy was used to obtain intermolecular distance restraints between human eIF1 and eIF5-CTD (Battiste and Wagner, 2000).

SAXS and ITC experiments are also described in more detail in the Extended Experimental Procedures. SAXS of eIF5, eIF1:eIF5-CTD, and eIF1:(eIF5-CTD+eIF2β-NTD) was measured in 20 mM Tris-HCl (pH 7.4), 200 mM NaCl, 0.5 mM TCEP, and eIF5-CTD at a final concentration of 90 μM, whereas eIF1 concentrations were in the range from 90 to 1,800 μM. Protein samples for ITC experiments were prepared in 20 mM Tris-HCl (pH 7.2), 200 mM NaCl, 0.5 mM TCEP, 0.5 mM EDTA. A MicroCal iTC_{200} calorimeter was run at an equilibrium temperature of 25°C. The concentration of the protein in the well was roughly ten times the estimated K_D, and the concentration of the protein in the syringe was seven times the one in the well.

Yeast Biochemistry and Genetics Experiments

Yeast biochemistry experiments were performed as described previously by Maag et al. (2006) and Nanda et al. (2009). eIF2-TC recruitment to 43S PICs experiments was carried out with a native gel assay as previously described by Acker et al. (2007). Unlabeled eIF1A was added to preassembled 43S·mRNA complexes (containing eIF1A-FL; fluorescently labeled), and anisotropy values were plotted as a function of time (Maag et al., 2006). GTP hydrolysis experiments were performed as described previously by Algire et al. (2005). Yeast genetics experiments were performed as described previously by Cheung et al. (2007), Nanda et al. (2009), Reibarkh et al. (2008), and Yamamoto et al. (2005). These experiments are written in more detail in the Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures, five tables, Extended Experimental Procedures, Extended Results, and an Extended Discussion and can be found with this article online at doi:10.1016/j.celrep.2012.04.007.

LICENSING INFORMATION

ACKNOWLEDGMENTS

This work was supported by NIH Grants CA68262 and GM47467 to G.W., NRSA-GM79970 to R.E.L., NIDDK-K01-DK085198 to H.A., G.W.'s GM47467 for L.E.L., GM64781 to K.A., NHBLI-T32HL07623-25 for P.R.H., and by the Intramural Research Program of the NIH (to P.M.-M. and A.G.H.). We thank T. Pestova for generously providing us with eIF1 mutant constructs. We would like to also thank L. Yang and M. Allier (Beamline X-9 NSLS, BNL) at Brookhaven National Laboratory. Use of the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. We thank E. Bedoya for purifying proteins and V. D'Souza and M. Durney for help with ITC data.

Received: October 17, 2011
Revised: February 16, 2012
Accepted: April 19, 2012
Published online: May 24, 2012

REFERENCES

Acker, M.G., Shin, B.S., Dever, T.E., and Lorsch, J.R. (2006). Interaction between eukaryotic initiation factors 1A and 5B is required for efficient ribosomal subunit joining. J. Biol. Chem. *281*, 8469–8475.

Acker, M.G., Kolitz, S.E., Mitchell, S.F., Nanda, J.S., and Lorsch, J.R. (2007). Reconstitution of yeast translation initiation. Methods Enzymol. *430*, 111–145.

Algire, M.A., Maag, D., and Lorsch, J.R. (2005). Pi release from eIF2, not GTP hydrolysis, is the step controlled by start-site selection during eukaryotic translation initiation. Mol. Cell *20*, 251–262.

Alone, P.V., and Dever, T.E. (2006). Direct binding of translation initiation factor eIF2gamma-G domain to its GTPase-activating and GDP-GTP exchange factors eIF5 and eIF2B epsilon. J. Biol. Chem. *281*, 12636–12644.

Asano, K., and Sachs, M.S. (2007). Translation factor control of ribosome conformation during start codon selection. Genes Dev. *21*, 1280–1287.

Asano, K., Clayton, J., Shalev, A., and Hinnebusch, A.G. (2000). A multifactor complex of eukaryotic initiation factors, eIF1, eIF2, eIF3, eIF5, and initiator tRNA(Met) is an important translation initiation intermediate in vivo. Genes Dev. *14*, 2534–2546.

Asano, K., Shalev, A., Phan, L., Nielsen, K., Clayton, J., Valásek, L., Donahue, T.F., and Hinnebusch, A.G. (2001). Multiple roles for the C-terminal domain of eIF5 in translation initiation complex assembly and GTPase activation. EMBO J. *20*, 2326–2337.

Battiste, J.L., and Wagner, G. (2000). Utilization of site-directed spin labeling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited nuclear overhauser effect data. Biochemistry *39*, 5355–5365.

Cheung, Y.N., Maag, D., Mitchell, S.F., Fekete, C.A., Algire, M.A., Takacs, J.E., Shirokikh, N., Pestova, T., Lorsch, J.R., and Hinnebusch, A.G. (2007). Dissociation of eIF1 from the 40S ribosomal subunit is a key step in start codon selection in vivo. Genes Dev. *21*, 1217–1230.

Conte, M.R., Kelly, G., Babon, J., Sanfelice, D., Youell, J., Smerdon, S.J., and Proud, C.G. (2006). Structure of the eukaryotic initiation factor (eIF) 5 reveals a fold common to several translation factors. Biochemistry *45*, 4550–4558.

Das, S., Maiti, T., Das, K., and Maitra, U. (1997). Specific interaction of eukaryotic translation initiation factor 5 (eIF5) with the beta-subunit of eIF2. J. Biol. Chem. *272*, 31712–31718.

Dominguez, C., Boelens, R., and Bonvin, A.M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. *125*, 1731–1737.

Fekete, C.A., Mitchell, S.F., Cherkasova, V.A., Applefield, D., Algire, M.A., Maag, D., Saini, A.K., Lorsch, J.R., and Hinnebusch, A.G. (2007). N- and C-terminal residues of eIF1A have opposing effects on the fidelity of start codon selection. EMBO J. *26*, 1602–1614.

Fletcher, C.M., Pestova, T.V., Hellen, C.U., and Wagner, G. (1999). Structure and interactions of the translation initiation factor eIF1. EMBO J. *18*, 2631–2637.

Hinnebusch, A.G. (2011). Molecular mechanism of scanning and start codon selection in eukaryotes. Microbiol. Mol. Biol. Rev. *75*, 434–467.

Huang, H.K., Yoon, H., Hannig, E.M., and Donahue, T.F. (1997). GTP hydrolysis controls stringent selection of the AUG start codon during translation initiation in *Saccharomyces cerevisiae*. Genes Dev. *11*, 2396–2413.

Hyberts, S.G., and Wagner, G. (2003). IBIS—a tool for automated sequential assignment of protein spectra from triple resonance experiments. J. Biomol. NMR *26*, 335–344.

Laurino, J.P., Thompson, G.M., Pacheco, E., and Castilho, B.A. (1999). The beta subunit of eukaryotic translation initiation factor 2 binds mRNA through the lysine repeats and a region comprising the C2-C2 motif. Mol. Cell. Biol. *19*, 173–181.

Lomakin, I.B., Kolupaeva, V.G., Marintchev, A., Wagner, G., and Pestova, T.V. (2003). Position of eukaryotic initiation factor eIF1 on the 40S ribosomal subunit determined by directed hydroxyl radical probing. Genes Dev. *17*, 2786–2797.

Maag, D., Algire, M.A., and Lorsch, J.R. (2006). Communication between eukaryotic translation initiation factors 5 and 1A within the ribosomal pre-initiation complex plays a role in start site selection. J. Mol. Biol. *356*, 724–737.

Maiti, T., and Maitra, U. (1997). Characterization of translation initiation factor 5 (eIF5) from *Saccharomyces cerevisiae*. Functional homology with mammalian eIF5 and the effect of depletion of eIF5 on protein synthesis in vivo and in vitro. J. Biol. Chem. *272*, 18333–18340.

Marintchev, A., Frueh, D., and Wagner, G. (2007). NMR methods for studying protein-protein interactions involved in translation initiation. Methods Enzymol. *430*, 283–331.

Martin-Marcos, P., Cheung, Y.N., and Hinnebusch, A.G. (2011). Functional elements in initiation factors 1, 1A and 2β discriminate against poor AUG context and non-AUG start codons. Mol. Cell. Biol. *31*, 4814–4831.

Nanda, J.S., Cheung, Y.N., Takacs, J.E., Martin-Marcos, P., Saini, A.K., Hinnebusch, A.G., and Lorsch, J.R. (2009). eIF1 controls multiple steps in start codon recognition during eukaryotic translation initiation. J. Mol. Biol. *394*, 268–285.

Passmore, L.A., Schmeing, T.M., Maag, D., Applefield, D.J., Acker, M.G., Algire, M.A., Lorsch, J.R., and Ramakrishnan, V. (2007). The eukaryotic translation initiation factors eIF1 and eIF1A induce an open conformation of the 40S ribosome. Mol. Cell *26*, 41–50.

Pestova, T.V., Lomakin, I.B., Lee, J.H., Choi, S.K., Dever, T.E., and Hellen, C.U. (2000). The joining of ribosomal subunits in eukaryotes requires eIF5B. Nature *403*, 332–335.

Pestova, T.V., and Kolupaeva, V.G. (2002). The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. Genes Dev. *16*, 2906–2922.

Rabl, J., Leibundgut, M., Ataide, S.F., Haag, A., and Ban, N. (2011). Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. Science *331*, 730–736.

Reibarkh, M., Yamamoto, Y., Singh, C.R., del Rio, F., Fahmy, A., Lee, B., Luna, R.E., Ii, M., Wagner, G., and Asano, K. (2008). Eukaryotic initiation factor (eIF) 1 carries two distinct eIF5-binding faces important for multifactor assembly and AUG selection. J. Biol. Chem. *283*, 1094–1103.

Saini, A.K., Nanda, J.S., Lorsch, J.R., and Hinnebusch, A.G. (2010). Regulatory elements in eIF1A control the fidelity of start codon selection by modulating tRNA(i)(Met) binding to the ribosome. Genes Dev. *24*, 97–110.

Singh, C.R., Curtis, C., Yamamoto, Y., Hall, N.S., Kruse, D.S., He, H., Hannig, E.M., and Asano, K. (2005). Eukaryotic translation initiation factor 5 is critical for integrity of the scanning preinitiation complex and accurate control of GCN4 translation. Mol. Cell. Biol. *25*, 5480–5491.

Sokabe, M., Fraser, C.S., and Hershey, J.W. (2012). The human translation initiation multi-factor complex promotes methionyl-tRNAi binding to the 40S ribosomal subunit. Nucleic Acids Res. *40*, 905–913.

Watanabe, R., Murai, M.J., Singh, C.R., Fox, S., Ii, M., and Asano, K. (2010). The eukaryotic initiation factor (eIF) 4G HEAT domain promotes translation re-initiation in yeast both dependent on and independent of eIF4A mRNA helicase. J. Biol. Chem. *285*, 21922–21933.

Yamamoto, Y., Singh, C.R., Marintchev, A., Hall, N.S., Hannig, E.M., Wagner, G., and Asano, K. (2005). The eukaryotic initiation factor (eIF) 5 HEAT domain mediates multifactor assembly and scanning with distinct interfaces to eIF1, eIF2, eIF3, and eIF4G. Proc. Natl. Acad. Sci. USA *102*, 16164–16169.

Yu, Y., Marintchev, A., Kolupaeva, V.G., Unbehaun, A., Veryasova, T., Lai, S.C., Hong, P., Wagner, G., Hellen, C.U., and Pestova, T.V. (2009). Position of eukaryotic translation initiation factor eIF1A on the 40S ribosomal subunit mapped by directed hydroxyl radical probing. Nucleic Acids Res. *37*, 5167–5182.

Capture of MicroRNA–Bound mRNAs Identifies the Tumor Suppressor miR-34a as a Regulator of Growth Factor Signaling

Ashish Lal[1,2,3☯]*, Marshall P. Thomas[1,2☯], Gabriel Altschuler[4☯], Francisco Navarro[1,2☯], Elizabeth O'Day[1,2], Xiao Ling Li[3], Carla Concepcion[5], Yoon-Chi Han[5], Jerome Thiery[1,2], Danielle K. Rajani[1,2], Aaron Deutsch[1,2], Oliver Hofmann[4], Andrea Ventura[5], Winston Hide[4], Judy Lieberman[1,2]*

1 Immune Disease Institute, Program in Cellular and Molecular Medicine, Children's Hospital Boston, Boston, Massachusetts, United States of America, **2** Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America, **3** Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **4** Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **5** Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, New York, United States of America

Abstract

A simple biochemical method to isolate mRNAs pulled down with a transfected, biotinylated microRNA was used to identify direct target genes of miR-34a, a tumor suppressor gene. The method reidentified most of the known miR-34a regulated genes expressed in K562 and HCT116 cancer cell lines. Transcripts for 982 genes were enriched in the pull-down with miR-34a in both cell lines. Despite this large number, validation experiments suggested that ~90% of the genes identified in both cell lines can be directly regulated by miR-34a. Thus miR-34a is capable of regulating hundreds of genes. The transcripts pulled down with miR-34a were highly enriched for their roles in growth factor signaling and cell cycle progression. These genes form a dense network of interacting gene products that regulate multiple signal transduction pathways that orchestrate the proliferative response to external growth stimuli. Multiple candidate miR-34a–regulated genes participate in RAS-RAF-MAPK signaling. Ectopic miR-34a expression reduced basal ERK and AKT phosphorylation and enhanced sensitivity to serum growth factor withdrawal, while cells genetically deficient in miR-34a were less sensitive. Fourteen new direct targets of miR-34a were experimentally validated, including genes that participate in growth factor signaling (*ARAF* and *PIK3R2*) as well as genes that regulate cell cycle progression at various phases of the cell cycle (cyclins D3 and G2, *MCM2* and *MCM5*, *PLK1* and *SMAD4*). Thus miR-34a tempers the proliferative and pro-survival effect of growth factor stimulation by interfering with growth factor signal transduction and downstream pathways required for cell division.

Citation: Lal A, Thomas MP, Altschuler G, Navarro F, O'Day E, et al. (2011) Capture of MicroRNA–Bound mRNAs Identifies the Tumor Suppressor miR-34a as a Regulator of Growth Factor Signaling. PLoS Genet 7(11): e1002363. doi:10.1371/journal.pgen.1002363

Editor: Michael T. McManus, University of California San Francisco, United States of America

Received July 18, 2011; **Accepted** September 13, 2011; **Published** November 10, 2011

Funding: This work was supported by a research grant (JL) and fellowship (FN) from the GSK-IDI Alliance and by NSF predoctoral fellowships (MPT, DKR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lieberman@idi.harvard.edu (JL); lalas@mail.nih.gov (AL)

☯ These authors contributed equally to this work.

Introduction

microRNAs (miRNAs) that promote cell differentiation, inhibit cell proliferation, or enhance DNA damage or stress-induced cell cycle arrest or death, and whose expression is reduced in some cancers, are candidate tumor suppressor genes [1]. One of the most well studied tumor suppressor miRNAs is miR-34a. Depending on cellular context [2], ectopic over-expression of miR-34a induces cell cycle arrest [3], senescence [4] or apoptosis [5]. miR-34a is up-regulated by p53 in response to DNA damage [6–8], but can also be transcriptionally activated independently of p53 [9,10]. miR-34a is located on chromosome 1p36, a locus deleted in neuroblastoma, breast, thyroid, and cervical cancer [11,12]. In other cancers, miR-34a expression is epigenetically reduced by hypermethylation [13]. miR-34a administration can inhibit tumor outgrowth in mice [4]. Thus miR-34a satisfies the criteria for a tumor suppressor gene.

The best way to understand the function of a miRNA is to identify the genes it regulates. In this study we sought to understand how miR-34a acts as a tumor suppressor by identifying its direct target genes. However, target gene identification is not straightforward because of the partial complementarity of the short ~22 nt miRNA sequence with the miRNA recognition element (MRE) of the target gene [14]. MRE pairing to the miRNA seed region (nt 2–7) contributes significantly to target gene recognition and is the basis for the most successful target gene prediction algorithms [15,16]. However, a perfect seed match is not necessary [17,18] and does not guarantee targeting [19]. miRNA target prediction algorithms typically predict hundreds to thousands of putative miRNA target genes, but most predicted target genes are not bona fide targets and the best algorithms sometimes miss key targets [17,19–21]. It is unclear how many target genes are in fact regulated by a given miRNA in any physiological context. Analysis of genes whose mRNA or protein expression decreases when a miRNA is overexpressed or increases when it is antagonized identifies genes that may be either direct targets or indirectly regulated [22]. Biochemical methods to

Author Summary

microRNAs (miRNAs) are small RNAs that regulate gene expression by binding to mRNAs bearing a partially complementary sequence. miRNAs decrease the stability or translation of mRNA targets, leading to reduced protein expression. Understanding the biological function of a miRNA requires identifying its targets. Here we developed a sensitive and specific biochemical method to identify candidate microRNA targets that are enriched by pull-down with a tagged, transfected microRNA mimic. The method was applied to miR-34a, a miRNA that inhibits cell proliferation. We found that miR-34a can potentially regulate hundreds of genes. Computational analysis of these genes suggested a novel function for miR-34a—suppression of the pro-proliferative response to diverse growth factors. This function complements the previously known role of miR-34a in blocking cell cycle progression. Thus, by reducing the expression of an extensive network of genes, miR-34a dampens growth factor signaling as well as its downstream consequences, promotion of cell survival and proliferation.

capture RNA-induced silencing complex (RISC)-bound mRNAs potentially provide a more direct way to identify miRNA-regulated target genes [23–25]. However, immunoprecipitation has mostly been used to define the general features of miRNA-regulated mRNAs and their MREs, rather than to identify the targets of a particular miRNA.

Already 36 putative miR-34a targets have been validated by luciferase reporter assays. These targets strongly support miR-34a's role as a tumor suppressor. They include genes that promote cell cycle progression through the G_1/S transition (*CCND1*, *CCNE2*, *CDK4*, *CDK6*, *MYC*, *MYCN* and *E2F3*) [3,8,10,12,26], enhance transcription (*MYB*, *HNF4A* and *FOXP1*) [9,27,28] or growth factor signaling (*MET*, *MEK1*, *AXL* and *RRAS*) [8,29–32], inhibit apoptosis (*BCL2*) [33] or p53 activity (*YY1*, *MTA2*, *SIRT1* and *MAGE-A*) [5,32,34,35], and promote stem cell survival (*NOTCH1*, *NOTCH2*, *LEF1 WNT1*, *DLL1*, *JAG1* and *CD44*) [32,36–40]. The diversity of direct miR-34a targets suggests that miR-34a acts pleiotropically by regulating many genes.

To identify additional direct target genes of miR-34a without bias and understand better how miR-34a functions, we optimized a simple biochemical method to isolate mRNAs that bind to transfected biotinylated (Bi-)miR-34a [41,42]. mRNAs significantly enriched in the Bi-miRNA pull-down with streptavidin relative to their cellular expression were candidate targets. The pull-down was performed in two unrelated cancer cell lines, K562 erythroleukemia cells and HCT116 colon carcinoma cells. p53 activates transcription of miR-34a [8]. Under basal conditions, p53-sufficient HCT116 cells highly express miR-34a, while p53-null K562 cells do not express it above background (data not shown). We selected disparate cell lines to identify genes that may be regulated in multiple cell types or more specifically in a particular context. Several thousand genes were significantly enriched in the miR-34a pull-down in each cell line and 982 were significantly enriched in both cell lines. Most known miR-34a target mRNAs expressed in these cells were pulled down with miR-34a. Despite the large number of genes significantly enriched in the miR-34a pull-down, 91% of a random list of 11 genes enriched in both cell lines contained miR-34a-regulated 3′UTR sequences. These results suggest that the pull-down is quite specific and that miR-34a potentially directly regulates hundreds of genes. Bioinformatic analysis of the pulled down genes or of genes down-regulated after miR-34a transfection suggested that miR-34a regulates a dense network of genes that transduce proliferative signals arising from growth factor stimulation. Multiple candidate target genes participate in RAS-RAF-MAPK signaling. In fact miR-34a knockout reduced sensitivity to growth factor withdrawal by serum starvation, while miR-34a transfection led to increased vulnerability. Fourteen novel miR-34a targets identified by the pull-down in both cell lines were experimentally verified, including *ARAF* and *PIK3R2* in the RAS-RAF-MAPK pathway, and additional target genes required for cell cycle progression, including cyclins D3 and G2, *MAD2L2*, *MCM2*, *MCM5* and *PLK1*.

Results

Isolation of mRNAs bound to a transfected biotinylated–miRNA

We modified a method [5] for capturing miRNA-mRNA complexes using streptavidin-coated beads from cells transfected with miR-34a biotinylated at the 3′-end of the mature strand. Control samples were transfected with a biotinylated *C. elegans* miRNA (Bi-cel-miR-67) (Figure 1A). Biotinylation did not interfere with miRNA-mediated gene suppression as measured by luciferase reporter assay (Figure 1B). Over-expressing Bi-miR-34a or miR-34a in K562 cells also similarly suppressed expression of known miR-34a target genes (Figure 1C). Moreover, immunoprecipitation of HA-tagged Ago1 or Ago2 in K562 cells cotransfected with Bi-miR-34a specifically enriched for miR-34a by ~4-fold and ~6-fold, respectively (Figure 1D). Thus the Bi-miRNA is incorporated into the RISC and functions like the unbiotinylated miRNA.

We next optimized conditions to capture known target gene mRNAs. In the Bi-miR-34a pull-down of K562 cells, known miR-34a target transcripts *CDK4* and *CDK6*, but not UBC (a housekeeping gene), were enriched 12 hr after transfection, and their capture plateaued at 24–48 hr (Figure 1E). Therefore, 24 hr was chosen for subsequent experiments. The specificity of the pull-down and applicability to other cell types was verified since *CDK4*, *CDK6* and *MYB* mRNAs were consistently enriched by transfection of Bi-miR-34a, but not Bi-cel-miR-67, in K562 (Figure 1F) and HCT116 (Figure S1A) cells. Streptavidin beads did not enrich for non-target *SDHA* and *UBC* mRNAs, and the specific target mRNAs were not pulled down in cells transfected with unbiotinylated miR-34a (data not shown). miR-34a was specifically enriched >40-fold in the Bi-miR-34a pull-down compared to the input lysate (Figure S1B). Modifications of the pull-down to include formaldehyde cross-linking and/or pre-isolation of RNAs in high molecular weight cellular fractions reduced the amount of captured RNA, but did not improve the relative enrichment for known target gene mRNAs (data not shown). To confirm that association of Bi-miRNAs with target mRNAs was not a post-lysis artifact, we performed streptavidin pull-downs after adding Bi-miR-34a or Bi-cel-miR-67 to cytoplasmic extracts of untransfected K562 cells. *CDK4*, *CDK6* and *MYB* mRNAs were not enriched when Bi-miR-34a was added post-lysis (Figure S1C). The general applicability of the pull-downs to enrich for miRNA target genes was also verified for another miRNA, miR-24 in HepG2 cells. Bi-miR-24 capture enriched for 3 known miR-24 targets (*H2AFX*, *E2F2* and *MYC* [43]) by 2–5-fold (Figure 1G).

Sensitivity of the Bi-miR-34a pull-down

We next used gene expression microarrays to identify putative miR-34a targets captured by Bi-miR-34a in duplicate experiments from K562 (p53 deficient) and HCT116 cells (p53 proficient)

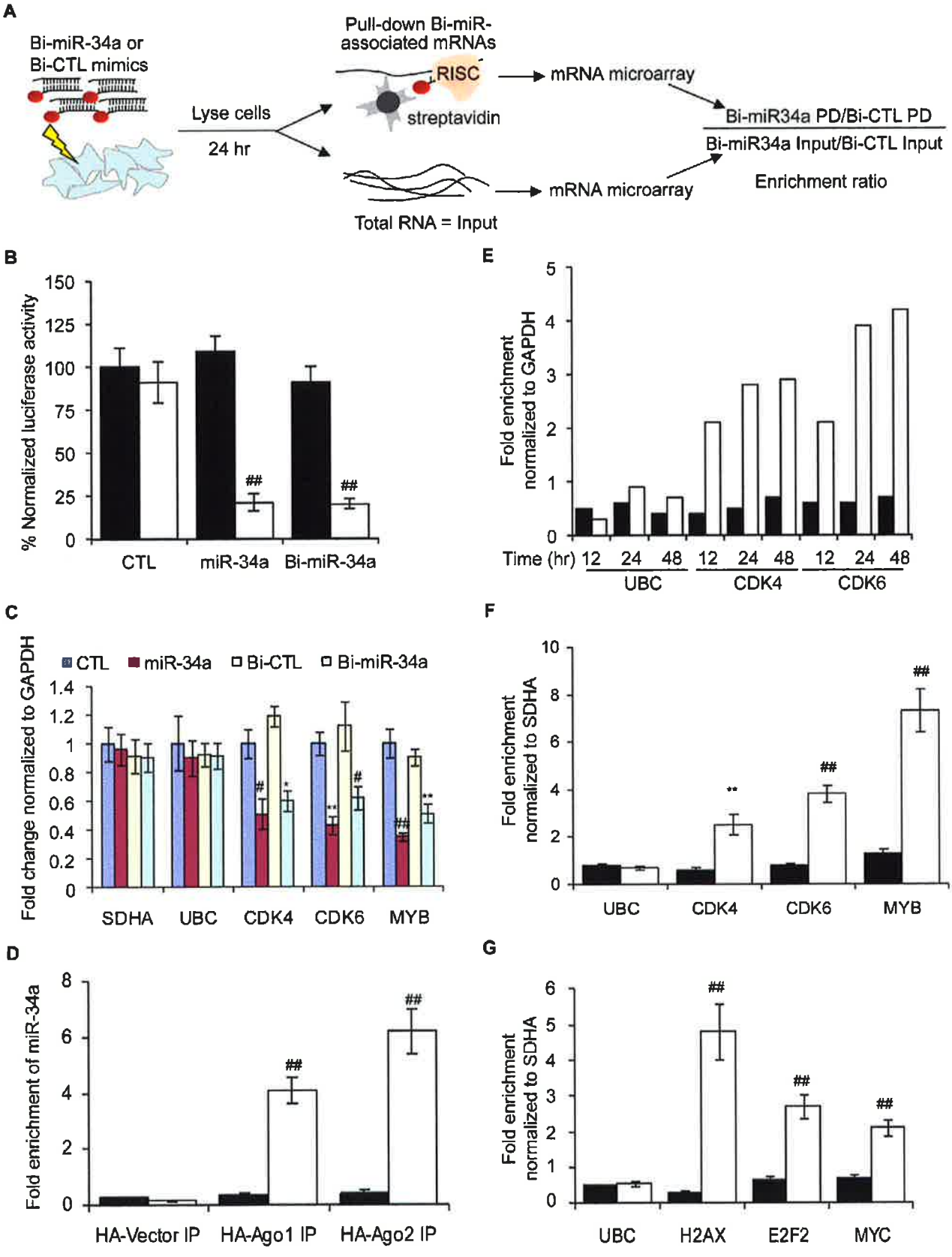


Figure 1. The Biotin-miRNA pulldown method. (A) Schematic of the Bi-miRNA pull-down (PD) assay. (B) Activity of 3′-biotinylated miR-34a (Bi-miR-34a) is similar to unbiotinylated miR-34a mimics by dual luciferase assay performed in HeLa cells cotransfected with psiCHECK-2 vector (black) or psi-CHECK-2 bearing a perfectly complementary sequence to miR-34a (psiCHECK-2-AS-miR-34a, white). Transfection with cel-miR-67 is the control (CTL). Luciferase expression was assayed after 24 hr; results are normalized to cells transfected with the luciferase vector and the CTL miRNA. (C) Bi-

miR-34a efficiently silences known miR-34a targets *CDK4*, *CDK6* and *MYB*. K562 cells were transfected with CTL miRNA, miR-34a, Bi-CTL or Bi-miR-34a mimics for 48 hr. Expression was measured by qRT-PCR normalized to *GAPDH*. The housekeeping genes *SDHA* and *UBC* are negative controls. (D) Cytoplasmic lysates from K562 cells were prepared 48 hr after cotransfection with Bi-CTL (black) miRNA or Bi-miR-34a (white) and a plasmid encoding HA-Ago1, HA-Ago2, or empty vector. Enrichment of miR-34a by HA immunoprecipitation was measured by qRT-PCR normalized to *U6*. Enrichment of Bi-miR-34a in the HA-immunoprecipitates suggests that Bi-miR-34a is incorporated into RISC. (E) Bi-miR-34a pull-downs optimally enrich targets 24 or 48 hr after transfection. K562 cells were transfected in duplicate with Bi-CTL (black) or Bi-miR-34a (white) mimics for the indicated times. Enrichment of known miR-34a targets (*CDK4* and *CDK6*) or control genes (*GAPDH* and *UBC*) was assessed by qRT-PCR relative to *GAPDH*. (F) The streptavidin pull-down enriches for miR-34a target genes in K562 cells transfected with Bi-CTL (black) or Bi-miR-34a (white) mimics. (G) Known miR-24 target mRNAs (*H2AX*, *E2F2* and *MYC*) are also pulled down with Bi-miR-24 in HepG2 cells reverse transfected 48 hr earlier with Bi-CTL (black) or Bi-miR-24 (white). Enrichment of target mRNAs in (F) and (G) was analyzed by qRT-PCR relative to *SDHA*. In all panels, data represent mean ± SD of 3 independent experiments. *, $p<0.05$, #, $p<0.01$, **, $p<0.005$, ##, $p<0.001$.
doi:10.1371/journal.pgen.1002363.g001

(Table S1). mRNA abundance in the streptavidin pull-down and input in Bi-miR-34a-transfected cells were separately normalized to their levels in Bi-cel-miR-67-transfected cells. For each biological replicate, the ratio of the abundance of the pull-down mRNA compared to the input mRNA for cells transfected with Bi-miR-34a versus Bi-cel-miR-67 was calculated, averaged and used to define the enrichment ratio {Bi-miR-34a PD/Bi-cel-miR-67 PD}/{Bi-miR-34a input/Bi-cel-miR-67 input}. Normalizing to the input improved identification of true targets in 2 ways – by reducing the background caused by highly abundant mRNAs that associate with streptavidin beads nonspecifically and by incorporating a measure of mRNA knockdown into the denominator of the ratio.

The miR-34a pull-downs enriched for 2416 genes in HCT116 cells (by ≥1 standard deviation (SD), enrichment ratio ≥2.5) and for 2816 genes in K562 cells (≥1 SD, enrichment ratio ≥3.3) (Figure 2A). The overlap of genes enriched ≥1 SD in both of these unrelated cell lines was 982 genes. To determine the sensitivity of the pull-down, we first looked at how many of the 36 published targets of miR-34a were captured in the K562 or HCT116 pull-downs (Figure 2B). Of the known expressed targets, 22 of 31 mRNAs (71%) were enriched in HCT116 cells and 14 of 29 (48%) were enriched in K562 cells. It should be noted that the choice of cut-off is somewhat arbitrary. Two additional known targets had enrichment ratios of 2.5–3.2 in K562 cells. The enrichment ratio ranged from 2.7–85. 12 genes were identified in both pull-downs. The enrichment ratio for the shared hits was not significantly different in K562 cells, which do not express miR-34a, compared to HCT116 cells, which do, suggesting that the pull-downs efficiently captured miR-34a targets even in cells that express endogenous miR-34a.

Analysis of genes down-regulated by miR-34a over-expression

To compare the mRNAs that associate with miR-34a to mRNAs that decrease with miR-34a over-expression, we measured mRNA abundance in cells transfected with miR-34a or cel-miR-67 by gene expression microarrays (Table S1). Genes whose mean mRNA level ratio decreased by at least 20% after miR-34a transfection were considered to be down-regulated either directly or indirectly by miR-34a. With this arbitrary cut-off (~1 SD), 2087 genes were down-regulated in HCT116 cells and 945 genes were down-regulated in K562 cells (Figure 2C). About a third of these transcripts in both cell lines were also pulled down with Bi-miR-34a (30% in HCT116, 36% in K562).

Down-regulated and pulled down mRNAs are enriched for miR-34a seed sequences

Many miRNA targets contain a perfect match to the miRNA seed region in their 3′UTR. We examined the frequency of 3′UTR matches to all hexamer sequences in miR-34a in the pull-down and down-regulated gene sets relative to all genes probed on the microarray (Figure S2A). Hexamer matches to nt 2–7 in the miR-34a seed region were significantly enriched in the pull-down (HCT116 p = 1.8E-95; K562 p = 2.4E-11) and down-regulated (HCT116 p = 1.7E-24; K562 p = 1.0E-11) datasets. There was also significant enrichment in the HCT116 pull-down genes for nt 13–19 exact matches, suggesting that base-pairing there enhances miRNA binding, as has previously been shown [44]. In both cell lines, seed enrichment was greater for the overlapping set of genes that was both pulled down and down-regulated by miR-34a. For genes in this overlap, exact matches to nt 2–7 were 1.8–2.0-fold more frequent per kb of 3′UTR than for all genes on the microarray. These data suggest that genes in the overlap may be more likely to be direct targets than genes identified by only one method or that a perfect seed match might enhance miRNA-mediated mRNA decay.

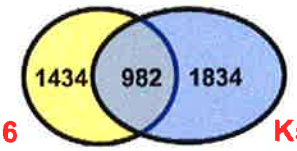
We next examined hexamer enrichment in the 982 genes enriched ≥1 SD in pull-downs from both HCT116 and K562 cells (Figure S2B). Seed matches were most enriched in the 3′UTRs of these genes, with the nt 2–7 match being the most abundant (1.7 fold more abundant than in all genes on the microarray (p = 8.4E-39). The coding region (CDS) of these genes also contained a highly significant enrichment for hexamer seed matches (p = 6.1E-13). These results are consistent with recent cross-linked RISC pull-downs that suggest that 25–50% of MREs may be in the CDS [23,25]. There was also a modest enrichment of hexamers matching the seed in the 5′UTR (p = 0.005). Thus the pull-down and down-regulated mRNAs were enriched for expected miRNA target sequence features.

We next analyzed whether mRNA expression of the enriched genes was reduced by miR-34a transfection in HCT116 cells (Figure 2D). The mRNAs of the 982 genes enriched in the miR-34a pull-down by ≥1 SD in both cell lines were significantly down-regulated after miR-34a transfection compared to the set of all genes expressed in the cell (p = 4.7E-80). The extent of down-regulation was comparable to the set of 469 TargetScan-predicted, evolutionarily conserved targets of miR-34a and significantly greater than in the larger list of 2904 poorly conserved, TargetScan-predicted genes (p = 1.6E-20). Increasing the cutoff for the enrichment ratio in the pull-down led to a greater proportion of highly down-regulated genes, indicating that a higher enrichment ratio correlates with more effective mRNA degradation and/or that highly enriched mRNAs are more likely to be miR-34a targets. Thus, the Bi-miR-34a pull-down enriches for known sequence and gene expression characteristics of *bona fide* miRNA targets.

Genes enriched in the miR-34a pull-down of both cell lines have a high probability of being direct miR-34a targets

To determine the specificity of the pull-down, we generated a random list (Table S2) of 11 genes enriched >2.5 fold in both pull-downs (median enrichment 3.5-fold, range 2.5–17.3). The random

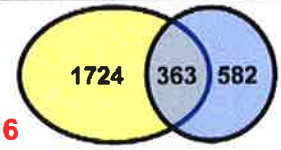
A



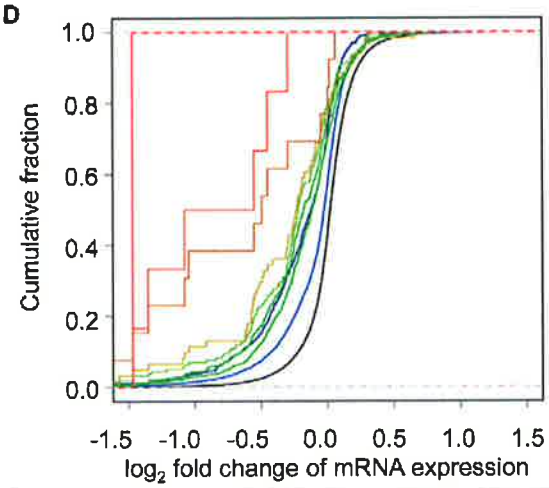
B

Target	Fold enrichment HCT 116	Fold enrichment K562	Ref.	Target	Fold enrichment HCT 116	Fold enrichment K562	Ref.
ACSL1	0.24	2.39	32	MAGEA2	Abs.	Abs.	35
ACSL4	**8.42**	**84.90**	32	MAGEA3	1.20	0.99	35
AXL	**16.97**	**18.32**	31	MAGEA6	1.27	0.95	35
BCL2	1.78	1.05	33	MAGEA12	Abs.	1.29	35
BIRC5	1.77	2.14	32	MEK1	**2.71**	1.92	29
CCND1	**3.66**	**40.29**	3	MET	**6.77**	0.53	8
CCNE2	**3.65**	**6.69**	8	MTA2	**8.37**	**11.22**	32
CD44	1.39	2.37	40	MYB	**8.94**	2.47	9
CDK4	**3.45**	2.73	8	MYC	1.47	**4.80**	10
CDK6	**22.80**	**62.52**	3	MYCN	Abs.	Abs.	26
DLL1	**8.30**	Abs.	39	NOTCH1	**7.13**	3.22	38
E2F3	**2.71**	**5.20**	12	NOTCH2	**7.18**	Abs.	38
EFNB1	**5.33**	**3.88**	32	RRAS	**63.86**	**32.42**	32
FOXP1	**16.98**	**6.69**	27	SIRT1	**9.84**	**10.22**	5
HNF4A	**8.35**	Abs.	28	TPD52	**13.31**	1.01	32
JAG1	0.99	Abs.	37	VCL	**6.20**	1.81	32
LDHA	1.43	1.17	32	WNT1	Abs.	Abs.	36
LEF1	Abs.	**32.65**	32	YY1	**4.83**	**52.82**	34

C

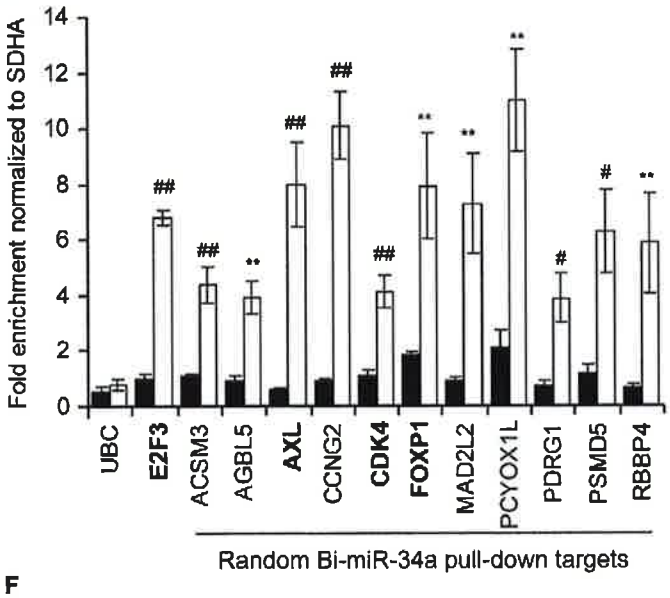


D

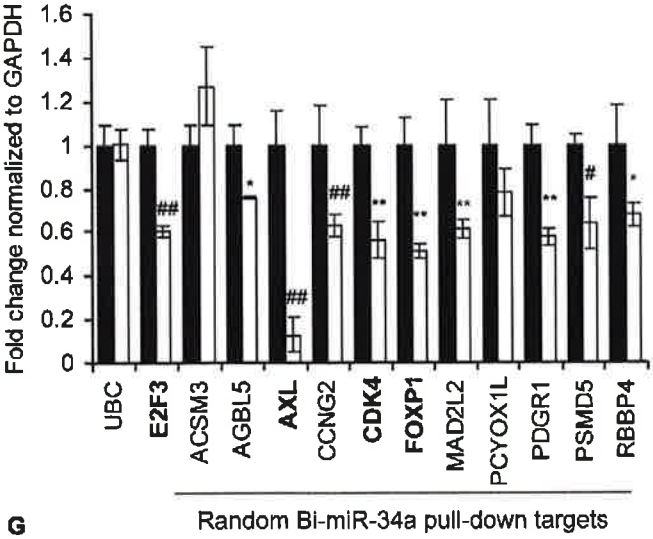


Genes		N	Genes		N
All genes	—	18,185	2.0 SD	—	167
TScan (all)	—	2904	2.5 SD	—	62
TScan (cons.)	—	469	3.0 SD	—	13
1.0 SD	—	982	3.5 SD	—	6
1.5 SD	—	408	4.0 SD	—	1

E



F



G

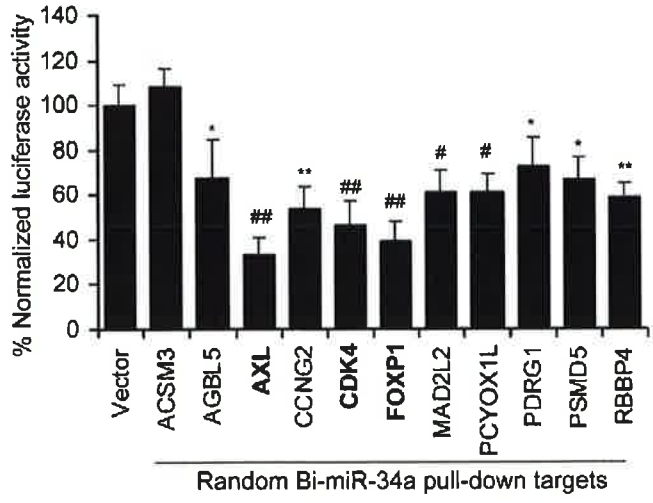


Figure 2. miR-34a pulls down transcripts of known and novel direct targets of miR-34a. (A) Overlap of genes enriched ≥1 SD in gene expression microarray analysis of Bi-miR-34a pull-downs from HCT116 and K562 cells. (B) Enrichment of previously described miR-34a target gene mRNAs in Bi-miR-34a pull-downs from HCT116 and K562 cells. Genes enriched ≥1 SD are indicated in bold. Abs = not expressed. (C) Genes down-regulated by ≥20% after miR-34a over-expression. (D) mRNA expression of candidate miR-34a targets identified by Bi-miR-34a pull-down in both HCT116 and K562 cells decreases after miR-34a over-expression. Cumulative distribution plots compare the extent of mRNA reduction of genes enriched to varying degrees in the Bi-miR-34a pull-down with conserved (cons) or all TargetScan (TScan)-predicted targets. Genes whose mRNAs are more highly enriched in the pull-down are progressively more likely to have reduced expression. (E) Pull-down of 11 mRNAs randomly selected from the set of genes enriched ≥2.5-fold by microarray in both HCT116 and K562 Bi-miR-34a pull-downs is confirmed by qRT-PCR relative to *SDHA* in K562 cells transfected for 24 hr with Bi-CTL (black) or Bi-miR-34a (white). *UBC* is a negative control and *E2F3* is a positive control. Three previously validated miR-34a targets (*AXL*, *CDK4* and *FOXP1*) in the random list of genes are indicated in bold. (F) miR-34a over-expression significantly decreases the expression of 9 of 11 of the randomly chosen candidate target genes. K562 cells were transfected with CTL miRNA (black) or miR-34a (white) mimics for 72 hr. Expression of random targets was measured by qRT-PCR normalized to *GAPDH*. Expression of 2 randomly selected genes (*ACSM3* and *PCYOX1L*) and the housekeeping mRNA *UBC* didn't change significantly. (G) miR-34a targets the 3′UTR of 10 of 11 randomly chosen targets. Luciferase activity was measured 48 hr after HeLa cells were cotransfected with the luciferase reporter psiCHECK2 bearing the 3′UTR of each gene and CTL miRNA or miR-34a mimics. Results obtained after miR-34a transfection were normalized to CTL miRNA. In (E–G), data represent mean ± SD of 3 independent experiments. *, $p<0.05$, #, $p<0.01$, **, $p<0.005$, ##, $p<0.001$.
doi:10.1371/journal.pgen.1002363.g002

list contained 3 known target genes (*AXL*, *CDK4* and *FOXP1*; *AXL* and *FOXP1* were not known when the list was generated). First, qRT-PCR analysis verified that the random gene mRNAs are pulled down by Bi-miR-34a and not Bi-cel-miR-67. All 11 mRNAs were enriched (~4–10 fold) by Bi-miR-34a pull-down in K562 cells, validating the microarray results (Figure 2E). miR-34a over-expression significantly down-regulated mRNA levels of 9 of 11 genes by 25–90% (Figure 2F). *PCYOX1L* expression declined by 20%, but the change was not significant. To test whether the 3′UTR of each gene could be regulated by miR-34a, the full 3′ UTR of each gene was cloned into a dual luciferase reporter plasmid. miR-34a repressed the 3′UTRs of 10 of 11 genes by ~20–80% (Figure 2G). Thus, miR-34a could regulate the 3′UTR of 91% of a random set of genes enriched in both miR-34a pull-downs. These results suggest that the Bi-miRNA pull-down is highly specific for identifying direct miRNA targets. An important implication of the large number of genes in the overlapping target list and the low false positive rate is that miR-34a is capable of regulating hundreds of genes.

miR-34a directly regulates growth factor signaling and cell cycle progression

To understand miR-34a's biological functions, we analyzed the cellular pathways whose genes were most enriched in the Bi-miR-34a pull-downs (Figure 3A). In both K562 and HCT116 cells, Bi-miR-34a pull-downs enriched for genes in pathways related to growth factor signaling and cell cycle control. Bi-miR-34a pull-downs enriched significantly for genes in the EGFR, TGF-β, interleukin, estrogen, and androgen receptor signaling pathways (Figure 3A). Many of these pathways utilize common downstream signaling molecules and have a well-established link to cancer. Genes in the MAPK pathway, activated by most growth factors, were highly enriched in the pull-downs for both cell lines. Growth factor signaling also activates cell proliferation. Genes involved in cell cycle regulation, especially the G_1/S transition, and the p53 response were enriched in both pull-downs, consistent with previously described targets and roles of miR-34a [3,4,7,8].

We performed a similar pathway enrichment analysis for genes down-regulated by miR-34a (Figure 3B), which includes both direct and indirect miR-34a targets. The downstream effects of growth factor signaling on cell proliferation and p53 activation were more prominent in the down-regulated genes than in the pulled-down gene set, especially in p53-sufficient HCT116 cells. Cell cycle and DNA repair pathways were enriched in genes down-regulated by miR-34a in both K562 and HCT116 cells. These results suggest that miR-34a directly inhibits growth factor signal transduction and cell cycle progression pathways, culminating in reduced expression of genes needed for cell proliferation.

A pathway enrichment analysis of the TargetScan-predicted targets of miR-34a (Figure S3) also highlighted the most significantly enriched pathways in the experimental pull-down and down-regulated gene sets, notably TGFβ and MAPK signaling and cell cycle and G_1/S transition. However, the significance of the enrichment was weaker and the strong role of miR-34a in growth factor signaling was less obvious.

miR-34a regulates a dense network of genes involved in signal transduction and cell cycle progression

To begin to understand regulation of growth factor signaling and cell proliferation at the gene level by miR-34a, an interactome of pulled down or down-regulated genes in HCT116 cells that participate in the significantly enriched pathways was generated (Figure 4). miR-34a potentially regulates the expression of critical genes involved in virtually every step and branch of growth factor signal transduction from ligand binding to downstream growth-promoting transcription factors. The putative direct targets included genes encoding multiple TGFβ and FGF isoforms, receptors for EGF, FGF, and insulin, and several oncogenic receptor tyrosine kinases, including *MET* and *AXL*. Several genes operating proximally in signal transduction, including *SRC*, *PLCG1* and *VAV2*, were selectively pulled down. miR-34a targets also included protein kinase subunits that activate downstream signaling, including subunits of protein kinase A and C. In the RAS-RAF-MAPK signal transduction pathway, putative directly regulated genes included *RRAS* and *RASA2*, *ARAF* and *BRAF*, *JAK2*, and 11 *MAPK* genes. Although knockdown of most of the targets would be expected to inhibit cellular activation by diverse growth factors, the genes also encode for some important inhibitors, including the ubiquitin ligase *CBLC*, *RASA2*, and 5 *DUSP* genes (MAPK phosphatases). The pull-down also captured 76 transcripts of transcription factors, including some that orchestrate the transcriptional response to signal transduction (including *STAT3*, *CREB1* and *CREB3*, *SP1*, *ELK1* and *SMAD4*).

A major downstream effect of growth factor signaling and its activated transcription factors is to stimulate cell proliferation. miR-34a is already known to suppress *E2F3* and some key cyclins and cyclin-dependent kinases that regulate the G_1/S transition. The miR-34a pull-down enriched for additional cyclins (*CCND3*, *CCNG2*), but also for transcripts of genes that inhibit the kinases that promote exit from G_1 (*CDKN1C* that encodes p57(KIP2), *CDKN2A* (p14(ARF)). Other enriched transcripts include *MCM5*, whose product is required to initiate DNA replication, and several genes required for mitosis (*PLK1*, *MAD2L2* and *CDC23*). Ectopic miR-34a expression led to down-regulation of mRNAs for many genes needed to replicate DNA, including 2 members of the initiating complex that assembles at origins of DNA replication, 7

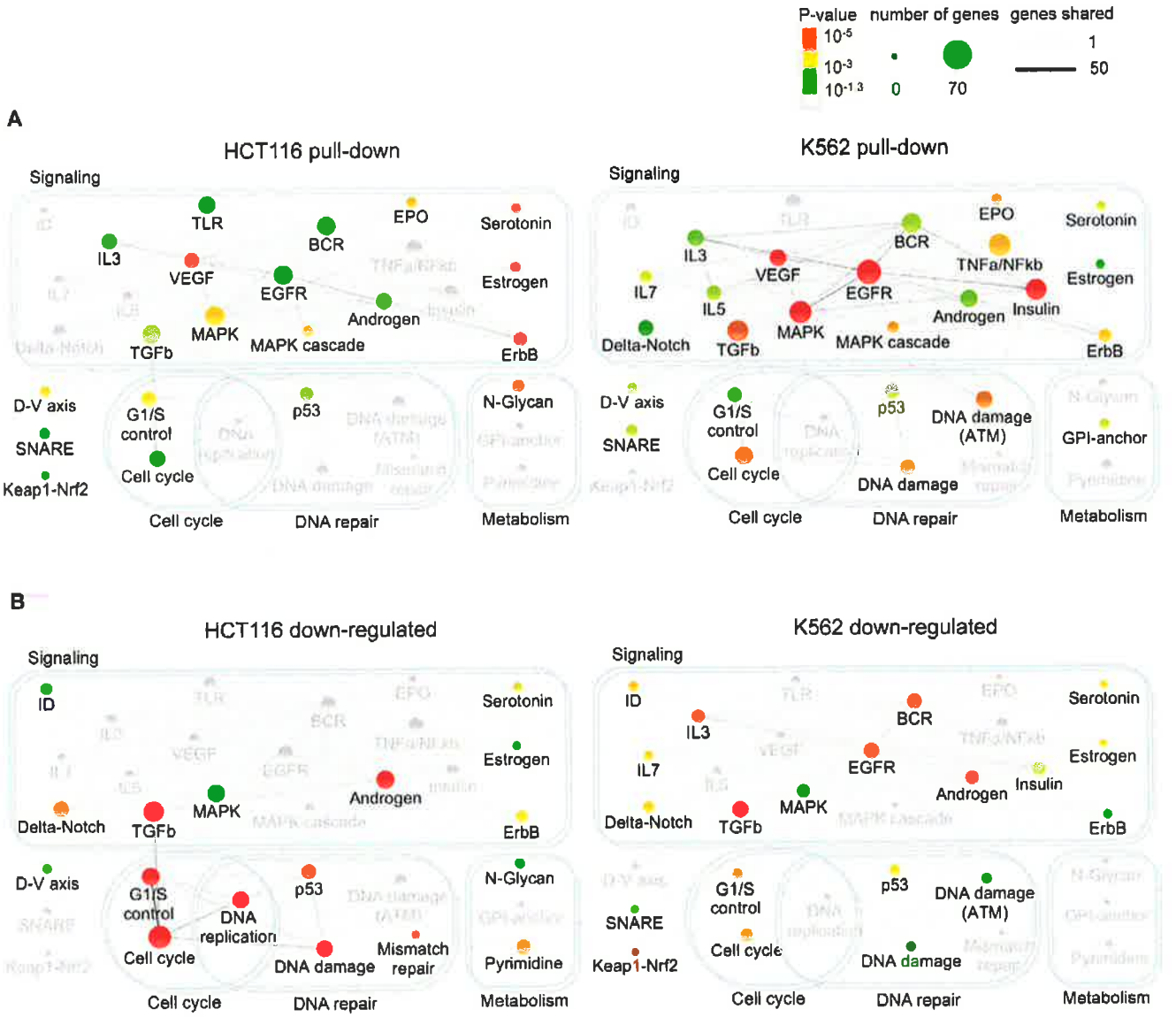


Figure 3. Genes in the Bi-miR-34a pull-down or down-regulated by miR-34a over-expression are enriched in growth factor signaling, cell cycle progression, and DNA repair pathways. Network of canonical pathways (Wikipathways and KEGG) significantly enriched for genes identified by Bi-miR-34a pull-down (A) or down-regulated following miR-34a over-expression (B) in HCT116 and K562 cells. Each pathway is represented by a node in the network. The node size increases with the number of identified genes in the pathway and the node color represents the p-value based on the hypergeometric distribution (see key). Pathways that are not significantly enriched in an experiment are still shown, but are in gray. The number of genes shared between two pathways is represented by an edge whose thickness increases with the number of shared genes. The pull-down enriched pathways (A) suggest that miR-34a extensively targets growth factor, signal transduction and cell cycle control pathways. The integrated outcome of both direct and indirect effects of miR-34a on gene expression in (B) is suppression of expression of genes participating in downstream signaling, cell cycle and DNA repair pathways.
doi:10.1371/journal.pgen.1002363.g003

components of the MCM complex, 4 DNA polymerases, and 5 components of the RFC complex, a cofactor for DNA polymerase. These results suggest that miR-34a not only interferes with the signaling that transduces the growth factor response, but also directly and indirectly suppresses the expression of numerous genes needed for cell proliferation.

miR-34a regulates cellular responses to growth factor signaling

The Ras–extracellular signal-regulated kinase (ERK) and phosphoinositide 3-kinase (PI3K)–AKT pathways are key transducers of the cellular response to growth factors. Since many candidate miR-34a target gene products act in pathways converging on ERK and AKT activation, we analyzed the effect of miR-34a over-expression on ERK and AKT phosphorylation. miR-34a transfection reduced basal phosphorylation of ERK and AKT in HCT116 and HeLa cells (Figure 5A, 5B), but not in A549 cells (Figure S4A). miR-34a over-expression both reduced basal proliferation in the absence of serum and blunted the ability of HCT116 (Figure 5C), HeLa (Figure 5D) and A549 (Figure S4B) cells to proliferate in response to serum growth factors. Conversely, immortalized mouse embryonic fibroblasts (MEFs) genetically deficient in miR-34a were more resistant to serum starvation than WT MEFs (Figure 5E). Apoptosis measured by annexin V and

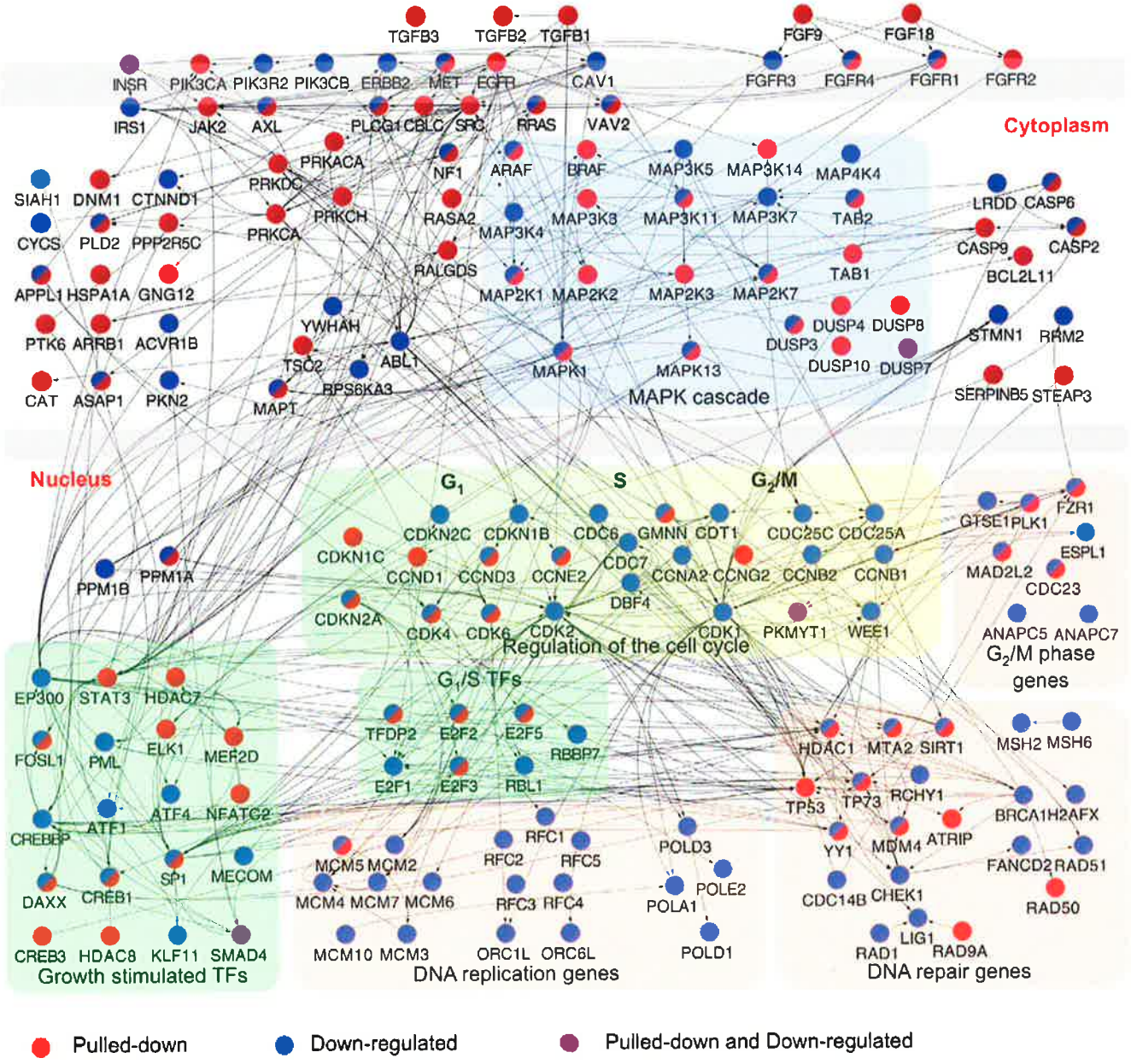


Figure 4. Interactome of genes in the enriched canonical pathways pulled down with Bi-miR-34a and/or down-regulated by miR-34a over-expression. Interactome of products of genes identified by Bi-miR-34a pull-down (red) or down-regulated by miR-34a over-expression (blue) in significantly enriched pathways (Figure 3) in HCT116 cells. Edges represent physical interactions. A dense network of genes involved in growth factor signaling and downstream effects on cell cycle progression and DNA repair is implicated.
doi:10.1371/journal.pgen.1002363.g004

propidium iodide staining was also significantly reduced in miR-34a$^{-/-}$ MEFs compared to wild-type MEFs after 24 hours of serum starvation (Figure 5F). Despite the strong difference in cell survival in cells deficient in miR-34a, expression of several known miR-34a targets did not differ significantly between wild-type and miR-34a$^{-/-}$ MEFs (data not shown). The lack of a notable difference may be due in part to compensatory up-regulation of miR-34b and miR-34c in miR-34a$^{-/-}$ MEFs (Figure 5G). These data suggest that miR-34a dampens the basal state of activation of proliferative and pro-survival pathways mediated by AKT and ERK by down-modulating multiple genes whose products contribute to their phosphorylation.

miR-34a directly targets genes that regulate ERK and AKT phosphorylation

To determine whether some of the candidate miR-34a target genes identified in the pull-down that participate in growth factor signaling are bona fide targets, we next tested miR-34a targeting of selected receptor-proximal (*AXL*, *MET* and *PIK3R2*) and more downstream (*ARAF* and *MEK1*) components of ERK and AKT signal transduction pathways. These 5 genes were both pulled down with Bi-miR-34a and down-regulated by miR-34a in HCT116 cells. ARAF is a serine/threonine protein kinase that phosphorylates and activates MEK1, which in turn phosphorylates ERK [45]. AXL is a receptor tyrosine kinase that stimulates cell

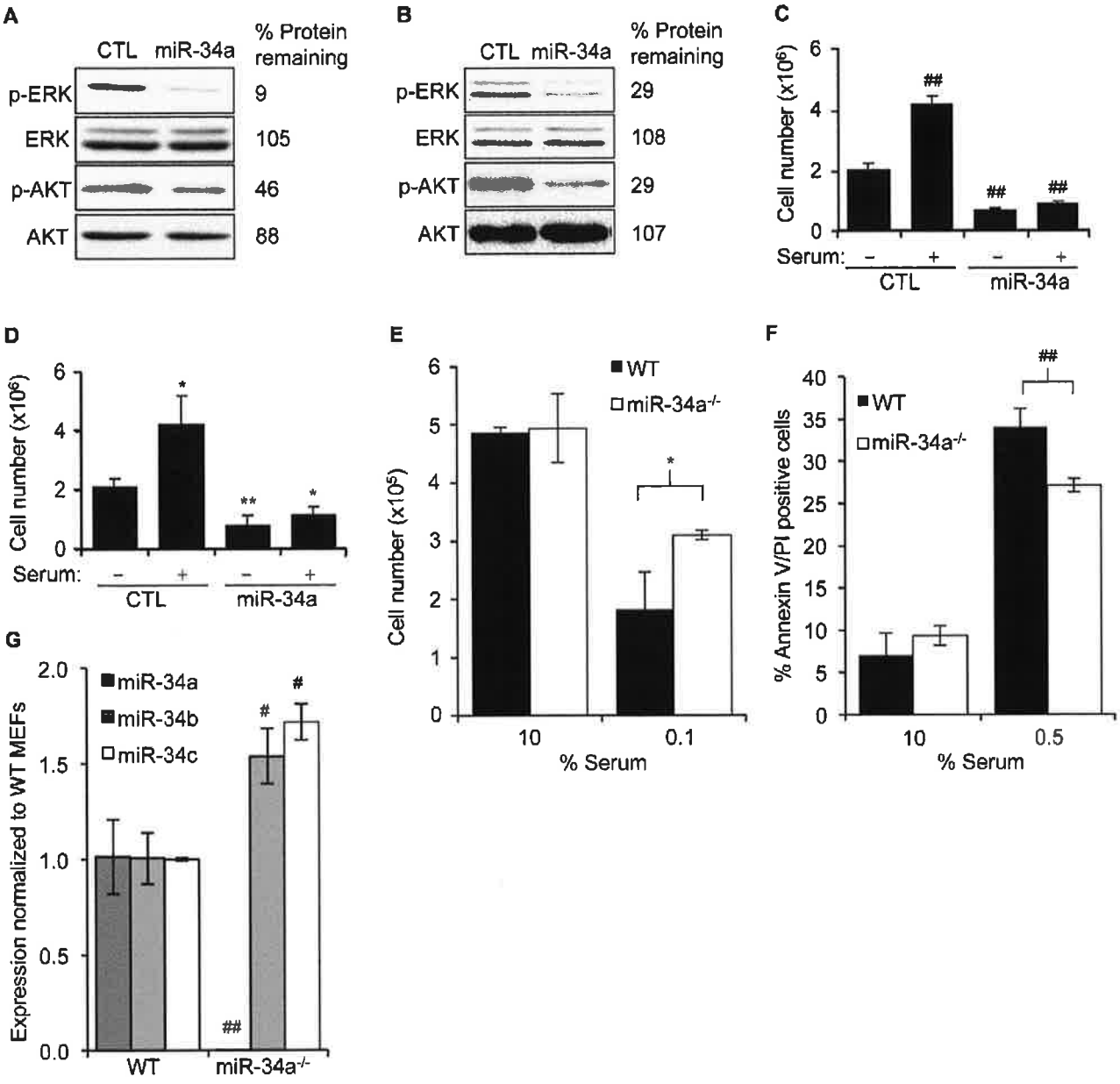


Figure 5. miR-34a expression suppresses cellular activation in response to serum growth factors. (A,B) miR-34a over-expression reduces basal phosphorylation of AKT and ERK as measured by immunoblot 48 hr after transfection of HCT116 (A) and HeLa (B) cells with control (CTL) miRNA or miR-34a mimics. Abundance of total ERK and AKT doesn't change. (C,D) miR-34a over-expression reduces cell proliferation in the absence of serum and suppresses the proliferative response of HCT116 (C) and HeLa (D) cells 24 hr after adding serum. (E) Total numbers of miR-34a$^{+/+}$ or miR-34a$^{-/-}$ MEFs after 24 hr of culture in 10% serum (10%) or 24 hr in 10% serum followed by 24 hr in 0.1% serum (0.1%). MEFs sufficient for miR-34a were more sensitive to serum starvation. (F) miR-34a$^{+/+}$ MEFs were more prone to apoptosis than miR-34a$^{-/-}$ MEFs after 24 hr of culture in reduced serum. (G) Expression of miR-34 family members in miR-34$^{+/+}$ (WT) and miR-34$^{-/-}$ MEFs assessed by qRT-PCR. miR-34a$^{-/-}$ MEFs expressed higher levels of miR-34b and miR-34c. In C,D,F, and G, data represent mean ± SD of 3 independent experiments. *, $p<0.05$, #, $p<0.01$, **, $p<0.005$, ##, $p<0.001$.
doi:10.1371/journal.pgen.1002363.g005

proliferation and also promotes metastasis [46,47]. PIK3R2 is a regulatory subunit of PI3K [48] and MET is a tyrosine kinase receptor that activates both PI3K and RAS [49]. *AXL*, *MET* and *MEK1* are described miR-34a targets [8,29,31], although *AXL* and *MEK1* were not known when these studies were performed.

The transcripts of all 5 genes were enriched 3–15-fold in the Bi-miR-34a pull-down by qRT-PCR, validating the microarray results (Figure 6A). Furthermore, over-expression of miR-34a down-regulated both the mRNA and protein levels of all 5 genes (Figure 6B, 6C). All but *ARAF* are also predicted miR-34a targets by TargetScan. To determine whether these genes are direct miR-34a targets, we tested the 3′UTRs for 4 of the genes (*ARAF*, *AXL*, *MEK1* and *MET*) by luciferase assay. miR-34a reduced reporter activity of these 3′UTRs by ~40–75% (Figure 6D). Using the PITA algorithm [50] to identify potential MREs in their 3′UTRs, we found 1 potential MRE in *AXL*, 2 in ARAF, 3 in *MEK1*, 4 in *PIK3R2* and 5 in *MET* (Figure S4). We tested repression of these MREs by miR-34a using luciferase assays. All 5 genes contained at least one miR-34a-responsive MRE (Figure 6E). Point mutations that disrupt the MRE-miR-34a interaction restored luciferase

activity, validating their regulation by miR-34a. Therefore, these 5 important genes in PI3K and MAPK signaling are all directly regulated by miR-34a.

miR-34a pull-downs identify new miR-34a targets that regulate cell cycle progression

Ectopic expression of miR-34a reduces expression of multiple direct target genes whose products facilitate the G_1/S transition (*CDK4*, *CDK6*, *CCND1*, *CCNE2* and *E2F3*). The pull-down identified novel genes acting at the G_1/S transition and genes involved in DNA replication and mitosis. Two cell cycle-regulating genes enriched in the miR-34a pull-down are in the random gene list and were already shown (Figure 2C–2E) to be miR-34a-regulated - *CCNG2*, which is most highly expressed in late S phase, and *MAD2L2*, a component of the mitotic spindle assembly checkpoint complex. To examine whether some of the other putative targets that participate in cell cycle progression are direct miR-34a targets, we focused on genes that were both pulled down and down-regulated by miR-34a in HCT116 cells (Table S2). Fourteen cell cycle-regulating genes (*CDK4*, *CDK6*, *CCNE2*, *E2F2*, *E2F3*, *E2F5*, *HDAC1*, *CDKN2A*, *MCM5*, *PKMYT1*, *PLK1*, *SMAD4*, *MAD2L2* and *CCND3*) met these criteria. Four of these (*CDK4*, *CDK6*, *CCNE2* and *E2F3*) are known miR-34a targets. We experimentally tested 5 of the 9 putative novel targets. These genes were *CCND3*, a cyclin that binds to CDK4 or CDK6 and regulates Rb phosphorylation; *MCM5*, a mini-chromosome maintenance (MCM) protein involved in initiating DNA replication, *MYT1*, a serine/threonine protein kinase that phosphorylates and inactivates CDC2, thereby negatively regulating cell cycle progression at the G_2/M transition; *PLK1*, a serine/threonine protein kinase required for mitotic spindle maturation; and *SMAD4*, a TGFβ-activated transcription factor that induces G_1 arrest and apoptosis. To determine whether these miR-34a pull-down genes are *bona fide* miR-34a target genes, we first verified that their transcripts associate with Bi-miR-34a (Figure 7A). After miR-34a over-expression, 3 of the 5 genes (*MCM5*, *PLK1* and *MYT1*) had reduced mRNA by at least 2-fold (Figure 7B) and all 5 had significantly reduced protein (Figure 7C). Two other MCM genes, *MCM2* and *MCM4*, also demonstrated a significant miR-34a-dependent reduction in mRNA, and their protein levels became undetectable in miR-34a-transfected cells.

To investigate whether these 5 genes are directly regulated, we measured changes in luciferase activity in HeLa cells after miR-34a co-transfection with reporters containing their 3′UTRs. The 3′UTRs of 4 of 5 of these genes (*CCND3*, *MCM5*, *PLK1* and *SMAD4*) were significantly repressed 30–60% by miR-34a (Figure 7D). The 3′UTR of *MYT1*, which bound to Bi-miR-34a and was down-regulated by miR-34a over-expression (Figure 7A, 7B), was not regulated by miR-34a. *MYT1* expression could be regulated by MREs outside the 3′UTR or indirectly. PITA and TargetScan were used to identify miR-34a MREs in the 3′UTRs of *CCND3*, *SMAD4*, *MCM5*, and *PLK1* (Figure 7E, Figure S5). *CCND3* MRE1, *SMAD4* MRE1 and *MCM5* MRE5 were significantly suppressed by miR-34a (Figure 7E, Figure S5). The *CCND3* and *SMAD4* MREs were predicted by TargetScan, while *MCM5* MRE5 contains a miR-34a hexamer seed match. Mutations that disrupt base pairing with miR-34a rescued luciferase expression, further confirming that these genes are direct miR-34a targets. Because the enrichment ratios for *MCM2* and *MCM4* in the pull-down (~2.3) were close to our cut-off, we also evaluated whether *MCM2* and *MCM4* might be direct targets. *MCM2* is a direct target as verified by mRNA enrichment in the pull-down, decrease in mRNA and protein following miR-34 over-expression, miR-34a regulation of its 3′UTR by luciferase activity and MRE identification (Figure 7A–7E). However, the *MCM4* 3′UTR was not active in luciferase assays. Collectively, these findings suggest that miR-34a acts as a master regulator of cell proliferation, directly suppressing many key genes that control cell cycle progression.

Discussion

Despite improvements in bioinformatic and experimental tools, distinguishing the direct targets of a miRNA from indirectly regulated genes remains challenging [14]. Here we describe a simple biochemical method to isolate candidate miRNA targets by streptavidin pull-down of mRNAs that associate with a transfected Bi-miRNA, and apply it to study miR-34a. Comparison of the set of mRNAs that directly associate with the Bi-miRNA with mRNAs down-regulated by miRNA over-expression makes it possible to distinguish the direct and indirect effects of a miRNA. Candidates identified by Bi-miR-34a pull-down have properties of validated miRNA targets: they are enriched for sequences complementary to the miR-34a seed and tend to decrease in expression with miR-34a over-expression. Genes that both decrease in mRNA abundance after over-expression and are isolated by Bi-miR-34a pull-down are further enriched for seed matches, indicating that either they are more likely true miR-34a targets or that a perfect seed match might enhance target mRNA degradation.

In our analysis we defined candidate direct targets using an arbitrary enrichment ratio cut-off of 1 SD, which corresponded to an enrichment of ≥2.5-fold for HCT116 cells and ≥3.3-fold for K562 cells. As the enrichment ratio cut-off was increased, mRNA suppression after ectopic miR-34a expression increased in tandem (Figure 2D). A more stringent cut-off would reduce the already low false positive rate, but also reduce the sensitivity to detect direct targets (Figure 2B). With this cut-off, we identify 71% of the known miR-34a targets expressed in HCT116 cells as "hits", but only 48% of the known expressed targets in K562 cells. If we had also chosen a 2.5-fold cut-off for K562 cells, our sensitivity for picking targets would have increased to 55%, while a 2-fold cut-off would have increased it to 69%. Since 10 of 11 genes in the random list of genes enriched by ≥2.5 fold by Bi-miR-34a pull-downs in both cells have 3′UTRs regulated directly by miR-34a by luciferase assay, a lower cut-off for the enrichment ratio might have increased sensitivity without an unacceptable false discovery rate. Some bona fide target genes are only enriched in the pull-down by ~2-fold; one of the novel genes we validated by identifying its MRE (*MCM2*) was only enriched by 2.3-fold in the pull-down of both cell lines. The low false positive rate of target identification demonstrated with the random gene list was also supported by the high degree of experimental validation of the growth factor signaling and cell cycle regulatory genes we chose to examine experimentally (Table S2). In all, we provided experimental evidence for 14 novel direct targets of miR-34a and identified 14 miR-34a MREs, of which 11 had a perfect hexamer seed match and the 3 others had perfect matches if G:U wobbles were allowed. Thus, the majority of genes we identified as regulated by miR-34a contain canonical 3′UTR MREs with good seed pairing. In the setting of over-expression by transfection, protein levels of all 11 genes we analyzed by immunoblot declined substantially. The few target genes that we tested for which we did not find miR-34a regulation of the 3′UTR might be false positives or might be direct targets, regulated by sequences in the 5′UTR or CDS. In fact we found enrichment for hexamer seed matches in these regions in the mRNAs pulled down with miR-34a, consistent with MRE

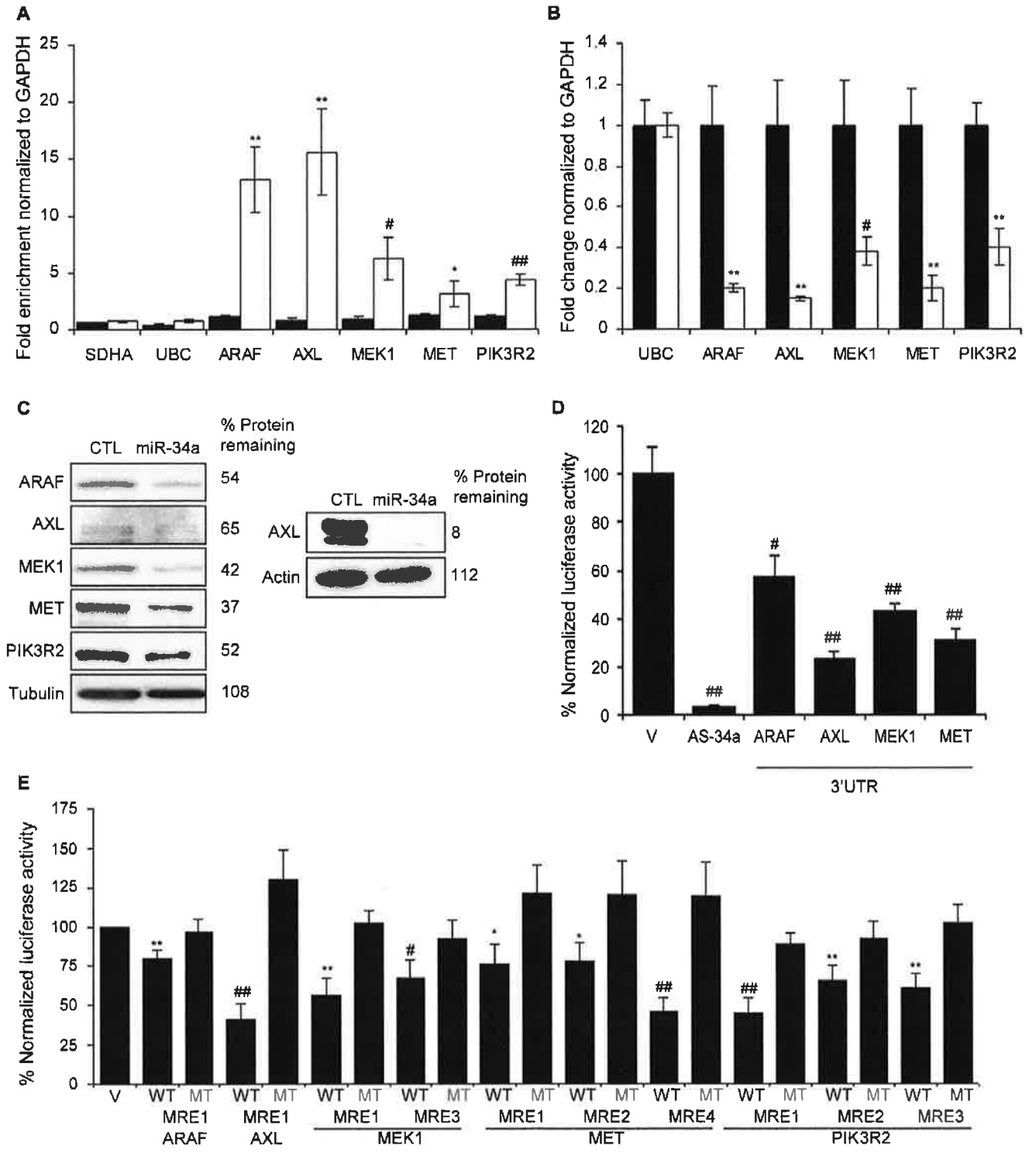


Figure 6. miR-34a directly inhibits growth factor signaling and signal transduction pathways by regulating novel genes. (A) Five genes involved in growth factor signaling or signal transduction are pulled down by Bi-miR-34a. HCT116 cells were transfected with Bi-cel-miR67 (CTL miRNA, black) or Bi-miR-34a (white) mimics for 24 hr and mRNA capture was measured by qRT-PCR normalized to *GAPDH*. Five of 5 (*ARAF*, *AXL*, *MEK1*, *MET* and *PIK3R2*) genes identified by microarrays, but not housekeeping mRNAs *SDHA* and *UBC*, are significantly enriched in the Bi-miR-34a pull-down. (B) miR-34a decreases *ARAF*, *AXL*, *MEK1*, *MET* and *PIK3R2* mRNAs, measured by qRT-PCR relative to *GAPDH*, in HCT116 cells transfected with CTL (black) or miR-34a (white) for 48 hr. *UBC* is a negative control gene. Relative mRNA levels were normalized to levels in CTL miRNA-transfected cells. (C) ARAF, AXL, MEK1, MET and PIK3R2 protein levels decline by immunoblot after miR-34a over-expression in HCT116 cells harvested 48 hrs after transfection with CTL miRNA or miR-34a mimic. Because of the low signal for AXL, an additional experiment probed for AXL with a longer exposure is shown at right. (D) miR-34a significantly regulates the 3′UTR of *ARAF*, *AXL*, *MEK1* and *MET* in HeLa cells co-transfected with a dual luciferase reporter bearing the 3′UTR of each gene and CTL miRNA or miR-34a for 48 hr. Insertion of a sequence fully complementary to miR-34a into the *Renilla luciferase* 3′UTR (AS-34a) is the positive control. Luciferase activity was normalized to results obtained with the empty vector (V). (E) Luciferase

reporters bearing PITA-predicted wild-type (WT) MREs from each target gene are significantly repressed in HeLa cells cotransfected with miR-34a. Point mutations (MT) that disrupt base pairing with miR-34a rescue reporter expression. MRE sequences are provided in Figure S4.
doi:10.1371/journal.pgen.1002363.g006

properties in recent cross-linking-RISC immunoprecipitation experiments [23,25].

Known targets may not have been identified by the pull-down for a variety of reasons. First, not all of the targets in the literature may be correctly assigned. Second, some known targets, such as CD44, are only modestly regulated by miR-34a [40]. The ratio that defines a "hit" is arbitrary. We set a relatively high threshold for identifying "hits" to maximize the specificity of the method (especially given the large numbers of enriched mRNAs in the pull-down), which came at the cost of sensitivity. Some known targets, which we did not designate hits with our 1 S.D. threshold of the enrichment ratio (which corresponded to >3.3 in K562 cells) had enrichment ratios of 2.5–3.2 in K562 cells. Other bona fide targets may have low, but detectable expression levels, and could have been missed due to the low sensitivity and inter-assay variability of microarray experiments. In addition to cellular variation in endogenous miRNA expression and RISC abundance, other context-dependent biological factors, such as target site accessibility, might vary due to the expression of RNA binding proteins, which could influence the efficiency of miRNA target site binding and the mechanism of targeting [51,52]. Cell-type specific expression of other MRE-containing genes that compete for miRNA binding could also influence the pull-down enrichment ratio [53]. Finally, some missed targets are likely to be false negatives.

Normalizing the pulled down mRNAs to their abundance in the input cellular mRNA was critical to eliminate from consideration highly abundant housekeeping mRNAs. Our pull-down method modified a previously developed protocol [41,42], which did not normalize the pull-down mRNAs to the input RNA. Many of the "hits" pulled down with Bi-miR-10a included ribosomal mRNAs, which may represent background binding of very abundant transcripts. Moreover, the miR-10a "hits" were not enriched for mRNAs containing miR-10a 3′UTR seed matches and were not down-regulated by miR-10a over-expression. In other work to be presented elsewhere, the pull-down method was used to identify genome-wide targets of miR-200c and miR-21. Importantly, the miR-200c and miR-21 pulled down mRNAs are also enriched for known targets and for 3′UTR seed sequences.

An advantage to the Bi-miRNA pull-down method described here is its simplicity. In contrast to mRNA expression-based target identification methods, Bi-miRNA pull-downs should identify only direct targets, excluding genes whose expression is indirectly modulated by changes in miRNA expression. Because the degree of mRNA suppression mediated by miRNAs is often small relative to changes in protein, methods that rely on changes in mRNA expression in response to manipulation of miRNA levels will necessarily miss some direct targets. Although the enrichment ratio takes into account a reduction in target gene mRNA in its denominator, the pull-down should not only identify target genes whose mRNA levels decline, but also those that are regulated primarily by inhibiting translation. Unlike approaches based on Ago pull-downs, the Bi-miRNA pull-down identifies the mRNAs directly associated with a specific miRNA, simplifying analysis of biological processes regulated by the miRNA.

The method described here without cross-linking does not directly identify MREs. The streptavidin pull-down method might, however, readily be modified to include cross-linking, RNase digestion of unbound mRNA segments and sequencing, similar to the HITS-CLIP protocol [23,24], to capture not only direct targets, but also identify MREs of an individual Bi-miRNA. Isolating RNAs associated with an individual miRNA rather than all RISC-associated RNAs in cells over-expressing the miRNA of interest might be a more direct way to define specific target sequences. Future bioinformatic studies of Bi-miRNA pull-down datasets could be used to better define in an unbiased manner the sequence features that dictate miRNA targeting, and could reveal non-canonical modes of targeting, such as those that contain only partial seed complementarity [17] or pairing to the central region of the miRNA [18] or that lie outside the 3′UTR. Indeed, in this work, we enriched for mRNAs with 5′UTR and CDS seed matches, indicating that some direct miR-34a targets may be regulated outside of their 3′UTR.

Only 29% of the 2416 enriched genes in the HCT116 pull-down had down-regulated mRNA levels by mRNA microarray analysis after over-expressing miR-34a for one day, while 10 of 11 randomly chosen genes in the pull-down had significantly decreased mRNA by qRT-PCR analyzed 72 hr after transfection. Thus although miRNAs may commonly lead to mRNA degradation, the degree of mRNA down-regulation of most genes is slight if cells are harvested within a day of transfection. mRNA microarrays may be too noisy to detect subtle changes in expression, unless the analysis is performed on many replicates. Our data also suggest that the kinetics of mRNA degradation may be slow. The early 24 hr time point used for the assay may have fortuitously enhanced our ability to capture miRNA-bound transcripts before too many had been degraded. Indirect effects of the miRNA are also likely to increase over time. The set of genes enriched in the miR-34a pull-down of both HCT116 and K562 cells contains 76 transcription factors or co-factors, whose suppression would reduce many mRNAs.

One important corollary of our results is that miR-34a likely directly regulates hundreds of genes. However, further experimental work is needed to assess how many of the hundreds to thousands of genes whose mRNAs associated with ectopic miR-34a are actually directly regulated by endogenous miR-34a. Possibly only a minority of potential targets is indeed directly regulated in an individual cell at any time. Based on our analysis (Figure 2D), the genes whose transcripts are most enriched in the pull-down may be the most significant targets in a given context. Additional experiments are needed to probe the functional consequences of miR-34a regulation of the genes we identified as targets. The directly regulated genes might vary considerably from cell type to cell type or even in the same cell lineage depending on differentiation state or environmental conditions. For this study we focused on the shared targets identified in two very different types of cells, rather than the ones that were unique to each cell-type. The pull-down method could be used in the future to compare miRNA target genes in different cellular contexts. Notably, the effect of miR-34a on cell signaling differed in the cancer cells we examined. Basal phosphorylation of AKT and ERK was reduced by miR-34a over-expression in HCT116 and HeLa cells (Figure 5), but not in A549 cells (Figure S4). Constitutively active RAS in A549 cells may override the effect of miR-34a in that context. Our results suggest that a dense network of genes that participate in common pathways, sometimes with opposing functions, is capable of being regulated by one miRNA. Although we observed a clear effect of genetic loss of miR-34a on the ability to cells to survive growth factor withdrawal, we did not see reduced expression in miR-34a$^{-/-}$ compared to wild-type

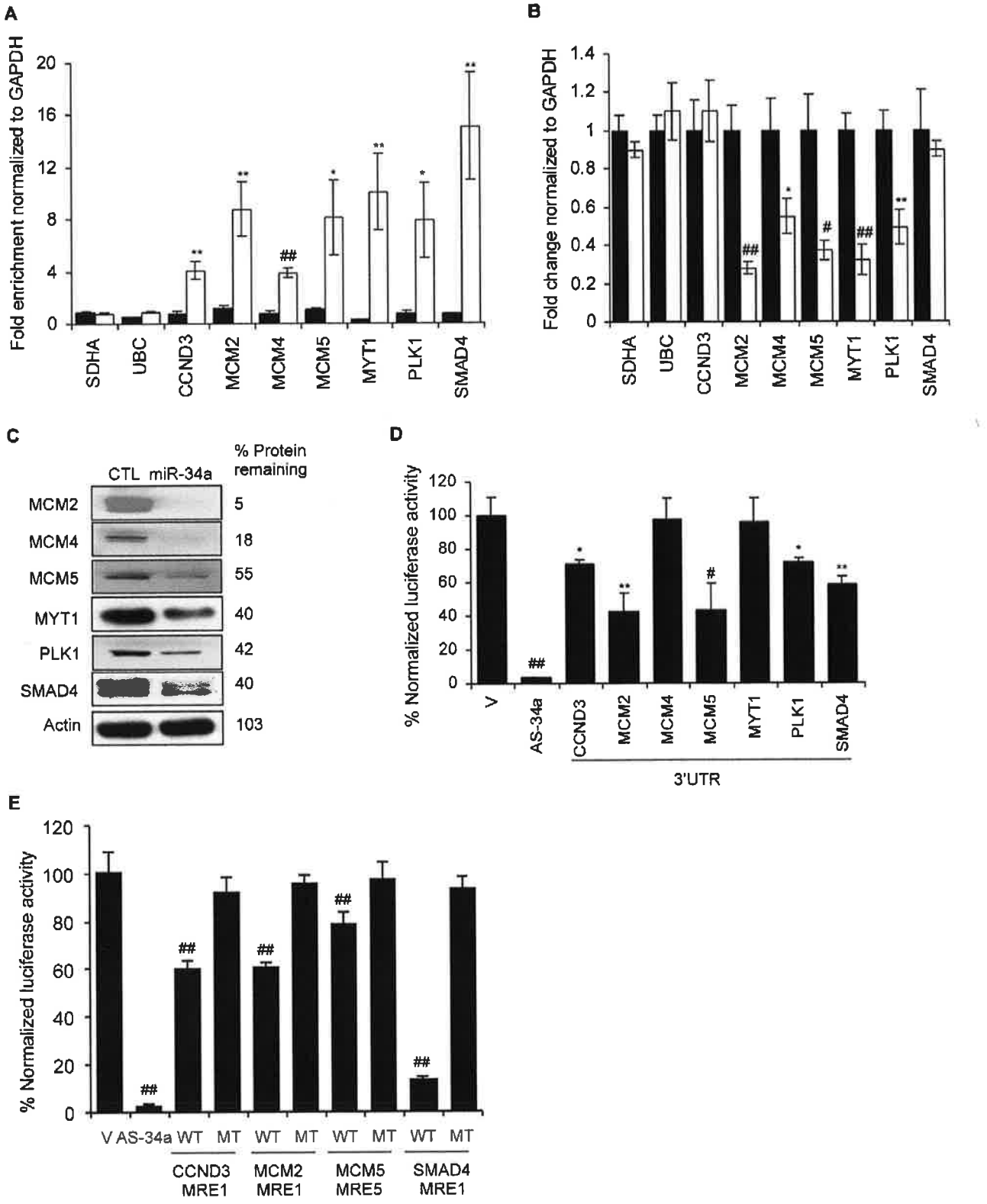


Figure 7. miR-34a pull-downs identify novel miR-34a targets involved in cell cycle progression. (A) Bi-miR-34a pull-down captures transcripts of cell cycle genes. Enrichment of candidate mRNAs in HCT116 cells transfected with Bi-cel-miR-67 miRNA (CTL, black) or Bi-miR-34a (white) for 24 hr was assessed by qRT-PCR analysis normalized to *GAPDH*. *SDHA* and *UBC* mRNAs are housekeeping genes not enriched in the pull-downs. (B) miR-34a over-expression significantly suppresses mRNA levels of 5 of 7 cell cycle genes tested. mRNA expression was analyzed by qRT-PCR relative to *GAPDH* performed on total RNA harvested from HCT116 cells transfected with CTL miRNA (black) or miR-34a (white) mimics for 48 hr. Two candidate target genes (*CCND3* and *SMAD4*) and the housekeeping genes *SDHA* and *UBC* are not significantly altered. (C) Protein levels of 6 of 6 cell cycle genes examined decrease with miR-34a over-expression. HCT116 cells were transfected with CTL miRNA or miR-34a mimics for 48 hr before immunoblot. β-Actin is a loading control. (D) miR-34a represses the 3′UTR of *CCND3*, *MCM2*, *MCM5*, *PLK1* and *SMAD4*, but not *MCM4* and *MYT1*. HeLa

cells were cotransfected for 48 hr with CTL miRNA or miR-34a mimics and psiCHECK-2 empty vector (V) or psiCHECK-2 containing the 3′UTR of each gene in the *Renilla luciferase* 3′UTR. The positive control reporter contained a perfectly complementary sequence to miR-34a (AS-34a). Relative luciferase activity in miR-34a-transfected cells is normalized to CTL miRNA-transfected cells. (E) Dual luciferase reporter plasmids bearing wild-type (WT) MREs from *CCND3*, *MCM2*, *MCM5* and *SMAD4* are significantly repressed by co-transfection with miR-34a in HeLa cells. Mutation of the seed region of each MRE (MT) rescues reporter expression. MRE sequences are provided in Figure S5. *, $p<0.05$, #, $p<0.01$, **, $p<0.005$, ##, $p<0.001$.
doi:10.1371/journal.pgen.1002363.g007

cells of some of the key miR-34a target genes we identified. Since growth factor signaling is so central to cell survival and proliferation, the permanent loss of miR-34a expression likely led to myriad compensatory changes. This seeming paradox supports the conclusions of our study – namely that a single miRNA may exert its biological effect by regulating expression of hundreds of genes. The capacity of miR-34a to potentially regulate so many genes that affect growth factor signaling may enable it to exert an effect in diverse contexts.

The numbers of genes that are actually regulated by miR-34a in any setting will likely depend on how strongly miR-34a is expressed. In our pull-down, we greatly over-expressed miR-34a. However, the level of over-expression throughout this study was not greater than endogenous miR-34a expression in some physiological settings, i.e. in K562 cells stimulated with phorbol ester where miR-34a increases 1000-fold [9]. There may be a target gene hierarchy – some genes regulated by low levels of miR-34a, others regulated only by high levels.

The dense network of cell signaling genes captured in the pull-downs suggests that an important function of miR-34a is to regulate the proliferative and activation responses to extracellular growth factors. Despite its function in regulating growth factor signaling and cell proliferation, we did not find a significant variation in miR-34a expression after serum starvation or when cells were synchronized in different phases of the cell cycle (data not shown). In this study we experimentally verified as direct miR-34a targets 5 growth factor signaling genes (*ARAF*, *AXL*, *MEK1*, *MET* and *PIK3R2*). miR-34a was previously shown to inhibit the G_1/S transition [3,8]. Here we identified 7 novel cell cycle-regulating direct targets that included genes also required for DNA replication and mitosis. The ultimate anti-proliferative effect of miR-34a integrates both direct consequences of suppressing expression of genes required for progression through the G1/S transition and at other steps of the cell cycle as well as indirect anti-proliferative effects from repressing the growth factor signaling pathways that activate cell cycle progression. Consistent with our genome-wide target gene analysis, miR-34a expression resets the basal state of ERK and AKT phosphorylation in several cell lines, rendering cells less responsive to growth factor signaling (Figure 5). This was shown both by miR-34a overexpression as well as by genetic deletion. miR-34a may reduce cellular sensitivity to growth factor signaling by suppressing many genes in multiple signal transduction pathways. miR-34a candidate targets include genes that are universally involved in transmitting growth factor activation signals as well as some that participate in specific pathways. The particular signaling genes that are suppressed in a given cell line will likely vary from cell to cell, depending on the growth factors to which the cell responds. These types of differences likely contribute to the incomplete overlap between the enriched pathways captured in the two hematopoietic and colon cancer cell lines examined here.

Materials and Methods

Cell lines

HCT116, K562, A549 and HeLa cells were from ATCC. miR-34a$^{+/+}$ and miR-34a$^{-/-}$ MEFs were generated from E14.5 littermate embryos. A full description of the mice will be published elsewhere. MEFs were transformed by infecting the cells with retroviruses encoding H-RAS-V12 and E1A and by selection with puromycin (1 μg/ml) and hygromycin (50 μg/ml). The plasmids for expression of H-RAS-V12 (plasmid 9051) and E1A (plasmid 18748) were obtained from Addgene. The VSV-G pseudotyped viruses were produced in 293T cells using the standard protocol. MEFs, HCT116, A549 and HeLa cells were grown in DMEM with 10% fetal bovine serum and supplemented with penicillin, streptomycin, HEPES, L-glutamine and β-mercaptoethanol, K562 cells were grown in RPMI containing 10% fetal bovine serum and the same supplements.

Transfection of miRNA mimics and plasmid DNA

For most experiments, 2×10^6 HCT116 or K562 cells were transfected with 200 pmol hsa-miR-34a or cel-miR-67 miRNA mimics (Dharmacon), using Amaxa nucleofection according to the manufacturer's protocol. Biotin was attached to the 3′-end of the active strand. HeLa and A549 cells were transfected with Lipofectamine 2000 and miRNA mimics at a final concentration of 50 nM (Invitrogen). To study the association of Bi-miRNAs with HA-Ago1 or HA-Ago2, pIRESNeo (Clontech) or pIRESNeo-HA-Ago1 or pIRESNeo-HA-Ago2 (Addgene) plasmids were co-transfected in six-well plates (2 μg/well, 1×10^6 cells/well) with 200 pmol Bi-miR-34a or Bi-cel-miR-67 using Amaxa as per the manufacturer's instructions.

RNA isolation and quantitative RT–PCR

Total RNA was isolated using Trizol reagent (Invitrogen), treated with DNase I (Ambion) and reverse transcribed using random hexamers and superscript III reverse transcriptase (Invitrogen). qRT-PCR was performed in triplicate samples using SYBR Green FastMix (Quanta) on a BioRad CFX96. mRNA levels were normalized to housekeeping genes *GAPDH*, *UBC* or *SDHA*. miRNA was quantified in triplicate using the TaqMan MicroRNA Assay (Applied Biosystems) as per the manufacturer's instructions and normalized to U6. Primer sequences are listed in Table S3.

Immunoblot

Whole cell lysates from transfected K562 or HCT116 cells were prepared using RIPA buffer. Proteins were analyzed by SDS-PAGE, transferred to nitrocellulose membranes and probed with the following antibodies: AXL [4566], ARAF [4432], MEK1 [9124], CDK4 [2906], MCM2 [3619], PKMYT1 [4282], PLK1 [4513], SMAD4 [9515], FOXP1 [2005], RBBP4 [4633], AKT [9272], pAKT ser-473 [4051], ERK [4370], pERK [9107] from Cell Signaling; MET [sc-161], MCM5 [sc-165995], E2F1 [sc-251], E2F3 [sc-879], CHEK1 [sc-8408] from Santa Cruz; ACSM3 [SAB1400253], MAD2L2 [SAB1400387], AGBL5 [AV53752], CCNG2 [AV03032], PSMD5 [WH0005711M1] from Sigma; MCM4 [06-1296] from Millipore; and PI3KR [610045], BD Biosciences. Western Blots were quantified by densitometry.

Biotin pull-down

HCT116 or K562 cells (1×10^6) were transfected in triplicate with Bi-miR-34a or Bi-cel-miR-67 (Dharmacon) as described

above and then cultured in six-well plates. Twenty-four hours later, the cells from 3 wells were pelleted at 500×g. After washing twice with PBS, cell pellets were resuspended in 0.7 ml lysis buffer (20 mM Tris (pH 7.5), 100 mM KCl, 5 mM $MgCl_2$, 0.3% NP-40, 50 U of RNase OUT (Invitrogen), complete mini-protease inhibitor cocktail (Roche Applied Science)), and incubated on ice for 5 min. The cytoplasmic lysate was isolated by centrifugation at 10,000×g for 10 min. Streptavidin-coated magnetic beads (Invitrogen) were blocked for 2 hr at 4°C in lysis buffer containing 1 mg/ml yeast tRNA and 1 mg/ml BSA (Ambion) and washed twice with 1 ml lysis buffer. Cytoplasmic lysate was added to the beads and incubated for 4 h at 4°C before the beads were washed five times with 1 ml lysis buffer. RNA bound to the beads (pull-down RNA) or from 10% of the extract (input RNA), was isolated using Trizol LS reagent (Invitrogen). The level of mRNA in the Bi-miR-34a or Bi-cel-miR-67 control pull-down was quantified by qRT-PCR or mRNA microarray. For qRT-PCR, mRNA levels were normalized to a housekeeping gene (*GAPDH*, *SDHA* or *UBC*). The enrichment ratio of the control-normalized pull-down RNA to the control-normalized input levels was then calculated.

Microarray analysis

Total RNA (independently in two experiments) was amplified, labeled and hybridized to Affymetrix U133 plus 2.0 mRNA microarrays. The quality of the RNA was assessed before performing the microarray and the quality of the microarray data was assessed using affyPLM and Affy software. The replicate data sets for the 4 sets of samples (pull-down and input for miR-34a and cel-miR-67) were compared using an unsupervised hierarchical clustering algorithm, which verified the similarity of the duplicates. The microarray data were normalized using RMA [7] to reduce interarray variation. The enrichment ratio {Bi-miR-34a PD/Bi-cel-miR-67 PD}/{Bi-miR-34a input/Bi-cel-miR-67 input} was calculated for each probe. For genes represented by multiple probes, the mean ratio for all the probes was calculated. Genes for which none of the probe hybridization signals exceeded the background were considered not expressed and were disregarded in the analysis. For informatic analysis of the PD data, genes whose enrichment ratio were ≥1 SD above background based on a log-normal distribution were considered "hits".

Gene down-regulation after miR-34a over-expression

HCT116 or K562 cells were transfected in independent duplicate experiments as above with unbiotinylated miR-34a or cel-miR-67 (Dharmacon) and total RNA was harvested 24 hr later and analyzed as above by gene expression microarrays. After normalization, fold changes for each probe were calculated as the ratio of input RNA from miR-34a-transfected cells to the ratio of input RNA from cel-miR-67-transfected cells. Genes were considered down-regulated if the ratio decreased by at least 20%, which corresponded to ~1 SD. To test the expression levels of putative target sets, each gene list was plotted in a cumulative distribution function (CDF) plot, and the Kolmogorov-Smirnov [KS] test was used for statistical comparisons between gene sets.

Analysis of miR-34a target genes by target prediction algorithms

To determine whether a gene was also a predicted target of miR-34a, the presence of miR-34a binding sites was analyzed using TargetScan 4.2 (http://www.targetscan.org/) [39,54,55] or PITA (http://132.77.150.113/pubs/mir07/mir07_prediction.html) [50].

Hexamer analysis

The mature hsa-miR-34a sequence was obtained from miRBase (http://mirbase.org/). All RefSeq human mRNA sequences were downloaded from NCBI in July 2009 (http://ftp.ncbi.nih.gov/). mRNAs were indexed by Entrez Gene ID; in cases where multiple sequences matched a gene ID, the sequence with the longest 3′UTR was selected. For each test gene list and miR-34a hexamer, the miR-34a hexamer frequency (hexamer matches per kb of sequence) was calculated. The frequency of hexamer matches for all genes on the microarray (the background set) was also determined. Gene IDs with no corresponding sequence in the database were excluded from analysis. Monte Carlo simulations of equally sized random gene sets (without replacement) were used to generate an empirical 2-tailed p-value for each gene set/hexamer combination. When $p<1E-4$, the p-value was calculated from curve fitting relative to the random background distribution.

Pathway enrichment analysis and network visualization

For each of the lists of down-regulated and pull-down-enriched genes, the p-value of over-representation in a suite of canonical pathways (KEGG [56] and Wikipathways [57]) was determined using the hypergeometric distribution. A visualization of the relationship between the enriched pathways ($p<0.001$) based on the number of overlapping genes was rendered using Cytoscape [58]. The network of gene-gene interactions underlying these relationships was constructed based on interactions supplied by MetaCore (GeneGo Inc). Physical, predicted and genetic interactions were used to connect the down-regulated and pull-down enriched genes within the significant signaling, cell cycle or DNA repair pathways. Signaling pathway genes with no connection to any other node were removed and the network was arranged according to predicted sub-cellular localization.

Luciferase assay

HeLa cells were cotransfected in 24 well plates using Lipofectamine 2000 (Invitrogen) with 50 nM miR-34a mimic or control miRNA mimic and 50 ng of psiCHECK2 (Promega) vector containing the MRE or 3′UTR of indicated genes cloned into the multiple cloning site of *Renilla* luciferase. After 48 hr of transfection (unless otherwise indicated) luciferase activities were measured using the Dual Luciferase Assay System (Promega) and Top count NXT microplate reader (Perkin Elmer) per manufacturer's instructions. All experiments were performed at least in triplicate. Results were normalized to those obtained in cells transfected with an empty vector. For some experiments, a perfectly complementary antisense sequence to the active strand of miR-34a was inserted into the multiple cloning site for use as a positive control. Data were normalized to *Firefly* luciferase and results from 3 independent experiments were compared. Sequence of primers used for cloning 3′UTRs for miR-34a target genes are listed in Table S4. MREs sequences were cloned into psiCHECK-2 by annealing complementary oligomers matching each MRE sequence (Figures S4, S5) with overhanging ends complementary to the XhoI and NotI sites of psiCHECK-2.

Cell growth experiments

HCT116, HeLa and A549 cells were transfected as described above. One day after transfection, cells were placed in serum-free medium or medium containing 10% fetal calf serum. 48 hours after the medium was changed, total cell numbers were counted. MEFs were plated at a density of 2.5×10^5 or 5×10^5 cells per well of a 6-well plate. The medium was changed to vary serum concentration 24 hr after plating. The MEFs were harvested 24 hr

later and counted using Trypan blue staining or stained in PBS+0.4% BSA with annexinV-APC (Invitrogen) at a 1:30 dilution, then washed once and stained with propidium iodide (4 μg/ml) (Sigma-Aldrich).

Supporting Information

Figure S1 (A) HCT116 cells were transfected with Bi-miR-34a or Bi-cel-miR-67 (CTL) and after 24 hr, abundance of known miR-34a target mRNAs (*CDK4*, *CDK6* and *MYB*) was measured by qRT-PCR analysis of pull-down RNA. *CDK4*, *CDK6* and *MYB* and not *UBC* (a housekeeping mRNA) were significantly enriched in the Bi-miR-34a pull-downs (white) and not the control pull-down (black). (B) K562 cells were transfected with Bi-miR-34a (white) or Bi-CTL (black), and RNA isolated from the streptavidin pull-down was analyzed by qRT-PCR for miR-34a and miR-24 (a control miRNA) after normalization to *U6*. miR-34a was ~50-fold higher in miR-34a pull-down as compared to control pull-down. miR-24 was not enriched and its levels were similar in each pull-down. (C) Addition of Bi-miR-34a (white) or Bi-CTL (black) to cytoplasmic extracts prepared from untransfected K562 cells does not enrich for known miR-34a target mRNAs, suggesting that the specific association of these mRNAs with Bi-miR-34a occurs in live cells and not post-lysis. Data in (B) are from 3 independent experiments and in (A) and (C) are from duplicate experiments.
(TIF)

Figure S2 Sequence characteristics of Bi-miR-34a pull-down targets. (A) Enrichment of hexamers matching each position of the mature miR-34a sequence in the HCT116 and K562 pull-down (red), down-regulated genes (blue), and genes down-regulated by miR-34a and pulled-down (yellow). Genes both enriched by Bi-miR-34a pull-down and down-regulated by miR-34a are the most enriched for miR-34a seed matches (B) Hexamer enrichment analysis for genes enriched in both HCT116 and K562 Bi-miR-34a pull-downs. Bi-miR-34a pull-down enriched for sequences matching two miRNA regions: the seed (positions 1–8) and a possible 3′ compensatory region (positions 13–19). Bi-miR-34a pull-down mRNAs are also enriched for CDS and 5′UTR matches to these sequences (*$p \leq 0.0001$).
(TIF)

Figure S3 Pathway networks representing the significant canonical pathways enriched for TargetScan conserved (A) and TargetScan non-conserved (B) target predictions.
(TIF)

Figure S4 miR-34a regulation of growth factor signaling. (A) Western blots of A549 cells transfected with miR-34a or CTL mimics. No reproducible change in pERK or pAKT was observed in these cells. (B) A549 cells were transfected with miR-34a or cel-miR-67 (CTL) mimics, and placed in normal growth medium with 10% serum (+) or growth medium lacking serum (−). Cells transfected with miR-34a did not proliferate in response to serum. Candidate miR-34a microRNA recognition elements (MRE) in the 3′UTR of *AXL*, *ARAF*, *MEK1*, *MET* and *PIK3R2* mRNAs predicted by PITA (see Materials and Methods). Numbers in parenthesis represent the location of the MRE in the 3′UTR. Wild-type MREs in (C) were repressed by miR-34a (see Figure 5F) whereas MREs that were not responsive to miR-34a are shown in (D). Point mutations that disrupt the base-pairing with miR-34a are shown in red in the mutant MREs.
(TIF)

Figure S5 Candidate miR-34a microRNA recognition elements (MRE) in the 3′UTR of *CCND3*, *MCM2*, *MCM5*, *PLK1* and *SMAD4* mRNAs predicted by PITA or TargetScan (see Materials and Methods). Numbers in parenthesis represent the location of the MRE in the 3′UTR (*PLK1* MRE2 spans the stop codon of *PLK1*). Wild-type MREs in (A) were repressed by miR-34a (see Figure 6E), whereas MREs that were not responsive to miR-34a are shown in (B). Point mutations that disrupt the base-pairing with miR-34a are shown in red in the mutant MREs.
(TIF)

Table S1 Genes enriched in Bi-miR-34a pull-downs or down-regulated by miR-34a over-expression in HCT116 and K562 cells.
(XLS)

Table S2 Experimental validation of miR-34a target genes.
(XLS)

Table S3 Sequence of primers used for qRT-PCR.
(XLS)

Table S4 Sequence of primers used for cloning 3′UTR of miR-34a target genes.
(XLS)

Acknowledgments

We thank Alex Amiet and Devin Leake (Dharmacon) for providing the Bi-miRNAs and A. Regev (Broad Institute), D. Chowdhury (DFCI), N. Caplen (NCI), P. Aplan (NCI) and Lieberman laboratory members for useful discussions.

Author Contributions

Conceived and designed the experiments: AL MPT GA FN AV WH JL. Performed the experiments: AL MPT FN EO XLL CC Y-CH JT DKR AD. Analyzed the data: AL MPT GA OH. Contributed reagents/materials/analysis tools: AV WH. Wrote the paper: AL MPT JL.

References

1. Garzon R, Calin GA, Croce CM (2009) MicroRNAs in Cancer. Annu Rev Med 60: 167–179. doi:10.1146/annurev.med.59.053006.104707.
2. He L, He X, Lowe SW, Hannon GJ (2007) microRNAs join the p53 network [mdash] another piece in the tumour-suppression puzzle. Nat Rev Cancer 7: 819–822. doi:10.1038/nrc2232.
3. Sun F, Fu H, Liu Q, Tie Y, Zhu J, et al. (2008) Downregulation of CCND1 and CDK6 by miR-34a induces cell cycle arrest. FEBS Letters 582: 1564–1568. doi:10.1016/j.febslet.2008.03.057.
4. Tazawa H, Tsuchiya N, Izumiya M, Nakagama H (2007) Tumor-suppressive miR-34a induces senescence-like growth arrest through modulation of the E2F pathway in human colon cancer cells. Proceedings of the National Academy of Sciences 104: 15472–15477. doi:10.1073/pnas.0707351104.
5. Yamakuchi M, Ferlito M, Lowenstein CJ (2008) miR-34a repression of SIRT1 regulates apoptosis. Proceedings of the National Academy of Sciences 105: 13421–13426. doi:10.1073/pnas.0801613105.
6. Raver-Shapira N, Marciano E, Meiri E, Spector Y, Rosenfeld N, et al. (2007) Transcriptional Activation of miR-34a Contributes to p53-Mediated Apoptosis. Molecular Cell 26: 731–743. doi:10.1016/j.molcel.2007.05.017.
7. Chang T-C, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, et al. (2007) Transactivation of miR-34a by p53 Broadly Influences Gene Expression and Promotes Apoptosis. Molecular Cell 26: 745–752. doi:10.1016/j.molcel.2007.05.010.
8. He L, He X, Lim LP, de Stanchina E, Xuan Z, et al. (2007) A microRNA component of the p53 tumour suppressor network. Nature 447: 1130–1134. doi:10.1038/nature05939.
9. Navarro F, Gutman D, Meire E, Caceres M, Rigoutsos I, et al. (2009) miR-34a contributes to megakaryocytic differentiation of K562 cells independently of p53. Blood 114: 2181–2192. doi:10.1182/blood-2009-02-205062.
10. Christoffersen NR, Shalgi R, Frankel LB, Leucci E, Lees M, et al. (2009) p53-independent upregulation of miR-34a during oncogene-induced senescence represses MYC. Cell Death Differ 17: 236–245.
11. Bagchi A, Mills AA (2008) The Quest for the 1p36 Tumor Suppressor. Cancer Research 68: 2551–2556. doi:10.1158/0008-5472.CAN-07-2095.
12. Welch C, Chen Y, Stallings RL (2007) MicroRNA-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells. Oncogene 26: 5017–5022.

13. Vogt M, Munding J, Grüner M, Liffers S-T, Verdoodt B, et al. (2011) Frequent concomitant inactivation of miR-34a and miR-34b/c by CpG methylation in colorectal, pancreatic, mammary, ovarian, urothelial, and renal cell carcinomas and soft tissue sarcomas. Virchows Arch 458: 313–322. doi:10.1007/s00428-010-1030-5.
14. Thomas M, Lieberman J, Lal A (2010) Desperately seeking microRNA targets. Nat Struct Mol Biol 17: 1169–1174. doi:10.1038/nsmb.1921.
15. Bartel DP (2009) MicroRNAs: Target Recognition and Regulatory Functions. Cell 136: 215–233. doi:10.1016/j.cell.2009.01.002.
16. Rajewsky N (2006) microRNA target predictions in animals. Nat Genet 38 Suppl: S8–13. doi:10.1038/ng1798.
17. Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, et al. (2009) miR-24 Inhibits Cell Proliferation by Targeting E2F2, MYC, and Other Cell-Cycle Genes via Binding to "Seedless" 3′UTR MicroRNA Recognition Elements. Molecular Cell 35: 610–625. doi:10.1016/j.molcel.2009.08.020.
18. Shin C, Nam J-W, Farh KK-H, Chiang HR, Shkumatava A, et al. (2010) Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing. Molecular Cell 38: 789–802. doi:10.1016/j.molcel.2010.06.005.
19. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, et al. (2008) The impact of microRNAs on protein output. Nature 455: 64–71. doi:10.1038/nature07242.
20. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics 25: 3049–3055. doi:10.1093/bioinformatics/btp565.
21. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature 455: 1124–1128. doi:10.1038/nature07299.
22. Johnson CD, Esquela-Kerscher A, Stefani G, Byrom M, Kelnar K, et al. (2007) The let-7 MicroRNA Represses Cell Proliferation Pathways in Human Cells. Cancer Res 67: 7713–7722. doi:10.1158/0008-5472.CAN-07-1083.
23. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature 460: 479–486. doi:10.1038/nature08170.
24. Zisoulis DG, Lovci MT, Wilbert ML, Hutt KR, Liang TY, et al. (2010) Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. Nat Struct Mol Biol 17: 173–179. doi:10.1038/nsmb.1745.
25. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. (2010) Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. Cell 141: 129–141. doi:10.1016/j.cell.2010.03.009.
26. Wei JS, Song YK, Durinck S, Chen Q-R, Cheuk ATC, et al. (2008) The MYCN oncogene is a direct target of miR-34a. Oncogene 27: 5204–5213. doi:10.1038/onc.2008.154.
27. Rao DS, O'Connell RM, Chaudhuri AA, Garcia-Flores Y, Geiger TL, et al. (2010) MicroRNA-34a Perturbs B Lymphocyte Development by Repressing the Forkhead Box Transcription Factor Foxp1. Immunity 33: 48–59. doi:10.1016/j.immuni.2010.06.013.
28. Takagi S, Nakajima M, Kida K, Yamaura Y, Fukami T, et al. (2010) MicroRNAs Regulate Human Hepatocyte Nuclear Factor 4α, Modulating the Expression of Metabolic Enzymes and Cell Cycle. Journal of Biological Chemistry 285: 4415–4422. doi:10.1074/jbc.M109.085431.
29. Ichimura A, Ruike Y, Terasawa K, Shimizu K, Tsujimoto G (2010) MicroRNA-34a Inhibits Cell Proliferation by Repressing Mitogen-Activated Protein Kinase Kinase 1 during Megakaryocytic Differentiation of K562 Cells. Molecular Pharmacology 77: 1016–1024. doi:10.1124/mol.109.063321.
30. Li N, Fu H, Tie Y, Hu Z, Kong W, et al. (2009) miR-34a inhibits migration and invasion by down-regulation of c-Met expression in human hepatocellular carcinoma cells. Cancer Letters 275: 44–53. doi:10.1016/j.canlet.2008.09.035.
31. Mudduluru G, Ceppi P, Kumarswamy R, Scagliotti GV, Papotti M, et al. (2011) Regulation of Axl receptor tyrosine kinase expression by miR-34a and miR-199a/b in solid cancer. Oncogene 30: 2888–2899.
32. Kaller M, Liffers S-T, Oeljeklaus S, Kuhlmann K, Röh S, et al. (2011) Genome-wide characterization of miR-34a induced changes in protein and mRNA expression by a combined pulsed SILAC and micro-array analysis. Molecular & Cellular Proteomics. Available: http://www.mcponline.org/content/early/2011/05/12/mcp.M111.010462.abstract. Accessed 1 Jul 2011.
33. Bommer GT, Gerin I, Feng Y, Kaczorowski AJ, Kuick R, et al. (2007) p53-mediated activation of miRNA34 candidate tumor-suppressor genes. Curr Biol 17: 1298–1307. doi:10.1016/j.cub.2007.06.068.
34. Chen Q-R, Yu L-R, Tsang P, Wei JS, Song YK, et al. (2011) Systematic Proteome Analysis Identifies Transcription Factor YY1 as a Direct Target of miR-34a. Journal of Proteome Research 10: 479–487. doi:10.1021/pr1006697.
35. Weeraratne SD, Amani V, Neiss A, Teider N, Scott DK, et al. (2011) miR-34a confers chemosensitivity through modulation of MAGE-A and p53 in medulloblastoma. Neuro-Oncology 13: 165–175. doi:10.1093/neuonc/noq179.
36. Hashimi ST, Fulcher JA, Chang MH, Gov L, Wang S, et al. (2009) MicroRNA profiling identifies miR-34a and miR-21 and their target genes JAG1 and WNT1 in the coordinate regulation of dendritic cell differentiation. Blood 114: 404–414. doi:10.1182/blood-2008-09-179150.
37. Pang RTK, Leung CON, Ye T-M, Liu W, Chiu PCN, et al. (2010) MicroRNA-34a suppresses invasion through downregulation of Notch1 and Jagged1 in cervical carcinoma and choriocarcinoma cells. Carcinogenesis 31: 1037–1044. doi:10.1093/carcin/bgq066.
38. Li Y, Guessous F, Zhang Y, DiPierro C, Kefas B, et al. (2009) MicroRNA-34a Inhibits Glioblastoma Growth by Targeting Multiple Oncogenes. Cancer Research 69: 7569–7576. doi:10.1158/0008-5472.CAN-09-0529.
39. Lewis BP, Shih I-hung, Jones-Rhoades MW, Bartel DP, Burge CB (2003) Prediction of Mammalian MicroRNA Targets. Cell 115: 787–798. doi:10.1016/S0092-8674(03)01018-3.
40. Liu C, Kelnar K, Liu B, Chen X, Calhoun-Davis T, et al. (2011) The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. Nat Med 17: 211–215. doi:10.1038/nm.2284.
41. Ørom UA, Lund AH (2007) Isolation of microRNA targets using biotinylated synthetic microRNAs. Methods 43: 162–165. doi:10.1016/j.ymeth.2007.04.007.
42. Ørom UA, Nielsen FC, Lund AH (2008) MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation. Molecular Cell 30: 460–471. doi:10.1016/j.molcel.2008.05.001.
43. Lal A, Pan Y, Navarro F, Dykxhoorn DM, Moreau L, et al. (2009) miR-24-mediated downregulation of H2AX suppresses DNA repair in terminally differentiated blood cells. Nat Struct Mol Biol 16: 492–498. doi:10.1038/nsmb.1589.
44. Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, et al. (2007) MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Molecular Cell 27: 91–105. doi:10.1016/j.molcel.2007.06.017.
45. Roberts PJ, Der CJ (2007) Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. Oncogene 26: 3291–3310. doi:10.1038/sj.onc.1210422.
46. Linger RMA, Keating AK, Earp HS, Graham DK (2008) TAM receptor tyrosine kinases: biologic functions, signaling, and potential therapeutic targeting in human cancer. Adv. Cancer Res 100: 35–83. doi:10.1016/S0065-230X(08)00002-X.
47. Gjerdrum C, Tiron C, Høiby T, Stefansson I, Haugen H, et al. (2010) Axl is an essential epithelial-to-mesenchymal transition-induced regulator of breast cancer metastasis and patient survival. Proceedings of the National Academy of Sciences 107: 1124–1129. doi:10.1073/pnas.0909333107.
48. Vanhaesebroeck B, Ali K, Bilancio A, Geering B, Foukas LC (2005) Signalling by PI3K isoforms: insights from gene-targeted mice. Trends in Biochemical Sciences 30: 194–204. doi:10.1016/j.tibs.2005.02.008.
49. Whittaker S, Marais R, Zhu AX (2010) The role of signaling pathways in the development and treatment of hepatocellular carcinoma. Oncogene 29: 4989–5005.
50. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. Nat Genet 39: 1278–1284. doi:10.1038/ng2135.
51. Elcheva I, Goswami S, Noubissi FK, Spiegelman VS (2009) CRD-BP Protects the Coding Region of [beta]TrCP1 mRNA from miR-183-Mediated Degradation. Molecular Cell 35: 240–246. doi:10.1016/j.molcel.2009.06.007.
52. Vasudevan S, Tong Y, Steitz JA (2007) Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. Science 318: 1931–1934. doi:10.1126/science.1149460.
53. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? Cell 146: 353–358. doi:16/j.cell.2011.07.014.
54. Lewis BP, Burge CB, Bartel DP (2005) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. Cell 120: 15–20. doi:10.1016/j.cell.2004.12.035.
55. Friedman RC, Farh KK-H, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19: 92–105. doi:10.1101/gr.082701.108.
56. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30. doi:10.1093/nar/28.1.27.
57. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) WikiPathways: Pathway Editing for the People. PLoS Biol 6: e184. doi:10.1371/journal.pbio.0060184.
58. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research 13: 2498–2504. doi:10.1101/gr.1239303.

HIV DNA is heavily uracilated, which protects it from autointegration

Nan Yan[a,b,1], Elizabeth O'Day[a], Lee Adam Wheeler[a], Alan Engelman[c], and Judy Lieberman[a,1]

[a]Immune Disease Institute and Program in Cellular and Molecular Medicine, Children's Hospital Boston, and Department of Pediatrics, Harvard Medical School, Boston, MA 02115; [b]Department of Internal Medicine and Department of Microbiology, University of Texas Southwestern Medical Center, Dallas, TX 75390; and [c]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Boston, MA 02115

Edited by John M. Coffin, Tufts University School of Medicine, Boston, MA, and approved April 22, 2011 (received for review February 23, 2011)

Human immune cells infected by HIV naturally contain high uracil content, and HIV reverse transcriptase (RT) does not distinguish between dUTP and dTTP. Many DNA viruses and retroviruses encode a dUTPase or uracil-DNA glycosylase (UNG) to counteract uracil incorporation. However, although HIV virions are thought to contain cellular UNG2, replication of HIV produced in cells lacking UNG activity does not appear to be impaired. Here we show that HIV reverse transcripts generated in primary human immune cells are heavily uracilated (>500 uracils per 10 kb HIV genome). We find that HIV DNA uracilation, rather than being dangerous, may promote the early phase of the viral life cycle. Shortly after reverse transcription, the ends of the HIV DNA are activated by the viral integrase (IN) in preparation for chromosomal insertion. However, the activated ends can attack the viral DNA itself in a suicidal side pathway, called autointegration. We find here that uracilation of target DNA inhibits the strand transfer of HIV DNA ends by IN, thereby inhibiting autointegration and facilitating chromosomal integration and viral replication. When uracilation is increased by incubating uracil-poor cells in the presence of increasing concentrations of dUTP or by infecting with virus that contains the cytosine deaminase APOBEC3G (A3G), the proportion of reverse transcripts that undergo suicidal autointegration decreases. Thus, HIV tolerates, or even benefits from, nonmutagenic uracil incorporation during reverse transcription in human immune cells.

After the HIV envelope fuses with a target cell plasma membrane to deliver the viral capsid into the cytoplasm, the viral genomic RNA is rapidly reverse transcribed into DNA. Soon thereafter within the HIV preintegration complex (PIC) in the host cell cytoplasm, HIV integrase (IN) activates the ends of the reverse transcripts by removing two nucleotides from the 3′ end of each strand through a process called 3′ processing to prepare for HIV DNA integration into chromosomal DNA. Three-prime processing makes the viral DNA vulnerable to autointegration, a suicidal side pathway in which the reactive ends attack viral DNA itself to produce deletion or inversion circles (1). Autointegration is a problem faced by retroviruses (2, 3) and mobile genetic elements, including bacteriophages and retrotransposons (4–6). Different retroviruses have evolved unique ways to control autointegration. Moloney murine leukemia virus employs the barrier-to-autointegration factor, which protects viral DNA through compaction to promote intermolecular integration (2, 7–9). We recently showed that the host SET complex, which contains three DNases (Ape1, Nm23-H1, and Trex1) and other DNA binding proteins, binds to the HIV PIC and protects it from autointegration (10). Knockdown of SET complex genes leads to increased HIV autointegration and correspondingly decreased chromosomal integration, resulting in ~2- to 10-fold less HIV replication.

The SET complex is a base excision repair (BER) complex—Ape1 is the rate-limiting BER endonuclease and the two other nucleases in the complex are postulated to have BER proofreading functions (11, 12). An important function of BER is to recognize and remove misincorporated uracils in DNA. The main target cells of HIV infection, CD4 T cells and macrophages, naturally contain unusually high cellular uracil content (e.g., 8.4% of the total nucleotide pool or 1/4.4 dUTP/dTTP in macrophages, 12.2% of the total nucleotide pool or 1/4.2 dUTP/dTTP in peripheral blood T lymphocytes, compared with 0.2% of the total nucleotide pool or 1/100 dUTP/dTTP in HeLa and many other mammalian cells) (13–15). The HIV reverse transcriptase (RT) lacks proofreading activity and does not efficiently distinguish dUTP from dTTP. As such, HIV DNA generated in infected human immune cells likely contains uracils from misincorporation, although how much is unknown. Uracils in HIV DNA can also occur through cytosine deamination by APOBEC3 proteins that are highly expressed in immune cells. The HIV accessory protein Vif binds and inhibits A3G and APOBEC3F (16). To replicate in a uracil-rich cellular environment, many viruses, including some retroviruses, encode a dUTPase, which limits cellular uracil concentrations and/or a uracil-DNA glycosylase (UNG) to remove uracil from viral DNA. Replication and maintenance of latency of dUTPase-deficient herpes simplex virus (HSV-1) are severely impaired in the murine nervous system where dUTP levels are high and cellular dUTPase is low (17). Inactivation of the *UNG* gene of vaccinia virus also compromises its viability (18, 19). Retroviruses that replicate via DNA intermediates are also subjected to the same threat by uracil. Some retroviruses, including β-retroviruses (Mason-Pfizer monkey virus, murine mammary tumor virus) and nonprimate lentiviruses (feline immunodeficiency virus, equine infectious anemia virus), encode a dUTPase. The retrovirus-encoded dUTPase is dispensable for replication in dividing cells where the cellular uracil:thymidine ratio is low, but is essential for replication in nondividing cells, such as macrophages (20, 21). Surprisingly, primate lentiviruses, such as HIV, encode neither dUTPase nor UNG, but still manage to replicate efficiently in nondividing cells with a uracil-rich environment. Previous studies suggested that host UNG2 is incorporated in HIV particles through interaction with viral Vpr or IN and hypothesized a role in removing uracils from HIV DNA (22, 23). However, it is unclear whether HIV benefits from packaging UNG2, because viruses generated from human *UNG2*$^{-/-}$ cells are as competent as those from WT cells, and suppressing UNG activity does not significantly affect HIV infection in macrophages (24).

To investigate the impact of uracil misincorporation on HIV replication and to understand how HIV manages to replicate in the uracil-rich cellular environment in human immune cells, we developed a way to quantify uracil residues in HIV reverse transcripts during infection of its natural target cells. We find that HIV DNA reverse transcribed in primary human immune CD4 T cells and macrophages is heavily uracilated (>500 uracils per 10 kb genome, or more than one in five thymidines from each strand are replaced with uracils). These results suggest that vi-

Author contributions: N.Y. and J.L. designed research; N.Y., E.O., and L.A.W. performed research; A.E. contributed new reagents/analytic tools; N.Y., E.O., A.E., and J.L. analyzed data; and N.Y., A.E., and J.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: nan.yan@utsouthwestern.edu or lieberman@idi.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1102943108/-/DCSupplemental.

rion-associated UNG2 does not effectively remove uracils from HIV DNA in infected human immune cells, supporting the notion that UNG2 may be dispensable for early stage HIV replication (24). We also found that uracilation of HIV DNA inhibits suicidal autointegration and increases viral infectivity using a single-round replication assay. Uracil misincorporation does not necessarily cause mutation per se because U:A pairs in chromosomal DNA are repaired faithfully by BER. Structural and enzymatic analysis of uracilated DNA and HIV IN suggests that uracil incorporation inhibits autointegration by inducing DNA distortions that block IN-mediated strand transfer, thereby favoring HIV DNA integration into chromosomal DNA in which uracil is efficiently removed. Thus, HIV tolerates, or may even benefit from, nonmutagenic uracil incorporation during reverse transcription in the uracil-rich environment of human immune cells.

Results

PCR Method to Estimate DNA Uracil Content. To estimate how many uracils are incorporated into HIV DNA in macrophages and CD4 T cells, we developed a PCR method. This method ("Taq/Pfu PCR") is based on the fact that the archaebacterial DNA polymerase Pfu is strongly inhibited by uracil-containing DNA, whereas Taq DNA polymerase, which lacks proofreading like HIV RT, is not (25, 26). A similar type of assay has been used to quantify uracil in DNA (27). To assess the differential sensitivity of these polymerases to uracil, we tested the ability of both polymerases to amplify synthetic oligonucleotides (100 nucleotides in length of HIV *gag* sequence) containing zero, one, and three uracils as target templates using a pair of primers on the ends (Fig. 1*A*). Taq was similarly efficient at amplifying all three templates, whereas Pfu only generated the expected PCR product with the uracil-free DNA template, confirming the inability of Pfu to replicate across uracils. To verify that the difference in detection of Taq and Pfu PCR products by agarose gel electrophoresis was not due to differences in polymerase activity or in detecting their products, we used SYBR green dye to track DNA synthesis in real time in both reactions and found that Taq and Pfu yielded similar cycle threshold (Ct) values with five randomly chosen primer pairs in the HIV genome and plasmid HIV DNA as template (Fig. 1 *B* and *C*). We then prepared cytosolic DNA extracts from HeLa–CD4 cells, which have a low endogenous ratio of dUTP/dTTP (15), after overnight incubation with increasing amounts of dUTP in the culture medium and infection with a single-round vesicular stomatitis virus glycoprotein (VSV-G)–pseudotyped HIV bearing a luciferase reporter gene (HIV-Luc) (10) for 10 h. This time was chosen because reverse transcription is largely completed, whereas HIV DNA remains mostly in the cytosol and chromosomal integration has generally not yet occurred (10). Cytosolic DNA was used as template for Taq and Pfu amplification with HIV gag primers. Taq yielded comparable amounts of HIV products regardless of dUTP supplementation, whereas Pfu PCR only yielded the expected 1-kb PCR product when cells were grown in the absence of added dUTP (Fig. 1*D*). The absence of a Pfu PCR product suggests that at least one uracil was incorporated in each strand of the HIV reverse transcript within the region examined. We obtained similar results using two other pairs of HIV gag primers.

HIV DNA Synthesized in Infected Human Immune Cells Contains Many Uracils. To estimate how many uracils are incorporated into HIV DNA in primary immune cells, we used the Taq/Pfu PCR assay to analyze cytosolic DNA from macrophages and CD4 T cells infected for 10 h with VSV-G pseudotyped or wild-type HIV (Fig. 2). Primer pairs were designed to amplify different size fragments from multiple locations throughout the HIV genome (Fig. 2*A*). Primer pairs amplifying longer regions are more likely to encounter a uracil (and thus block Pfu amplification) than primer pairs expanding shorter regions. The percentage of fragments of a given length that are "uracil positive" (amplified by Taq, but not Pfu), should provide a lower limit estimate of the

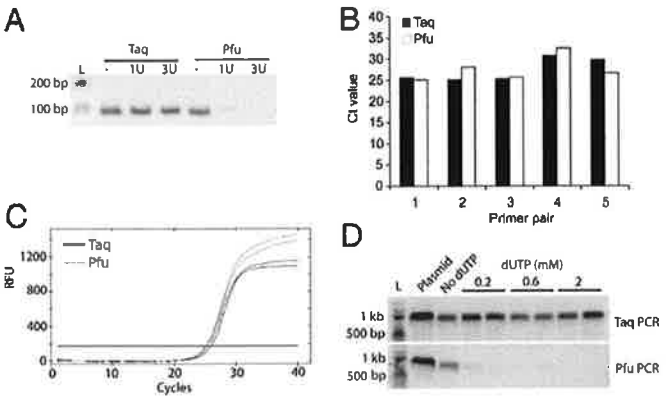


Fig. 1. Taq/Pfu PCR method to assess DNA uracil content (*A*) DNA gel electrophoresis of Taq/Pfu PCR products using synthetic 100 nt DNA oligonucleotides, containing 0, 1, or 3 uracils as templates. Oligonucleotide sequences are provided in Table S1. L, DNA ladder (same below). (*B* and *C*) Real-time analysis of Taq/Pfu PCR. SYBR-green real-time PCR was carried out with DNA polymerase Taq or Pfu using five primer pairs randomly chosen from within the HIV genome and HIV-GFP plasmid DNA as a template. Ct value for each qPCR is shown in *B*. A representative qPCR trace plot (primer pair 1, in duplicate) is shown in *C*. (*D*) Taq/Pfu PCR assay using DNA isolated from infected HeLa–CD4 cells cultured with indicated amounts of dUTP supplementation. HIV-GFP plasmid was used as a control ("plasmid"). *A* and *D* show representative gel images from at least three independent experiments.

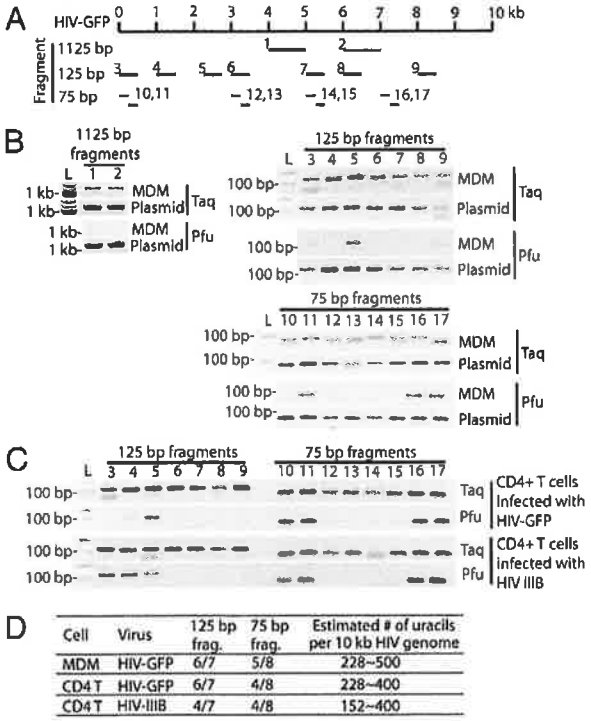


Fig. 2. HIV DNA is heavily uracilated in infected human immune cells. (*A*) Schematic showing the locations of primer pairs (numbered) in the HIV-GFP genome used for uracil mapping. The numbers in this diagram correspond to labels for the gel images in *B* and *C*. (*B*) Taq/Pfu PCR using DNA isolated from HIV-GFP–infected MDM or plasmid DNA (as a control). (*C*) Taq/Pfu PCR using DNA isolated from CD4 T cells infected with HIV-GFP or HIV_{IIIB} virus. (*D*) A table summarizing the findings from *B* and *C*. All primer pairs yielded expected PCR products in Taq-PCR or PCR reactions using plasmid as a template. Primer pairs that failed to yield a PCR product in Pfu-PCR were scored as uracil positive for that fragment. See text for details on how the estimated number of uracils per 10 kb HIV genome was calculated.

MICROBIOLOGY

number of total uracils incorporated into full length ~10-kb HIV DNA. A "uracil positive" fragment should contain at least 2 uracils, one in each strand. Primer pairs (25 nucleotides long) were designed to examine regions of three sizes (1,075, 75, and 25 bp) to produce PCR fragments of 1,125, 125, and 75 bp, respectively. An array of Taq/Pfu PCR reactions was performed to determine which fragments are amplified by Taq, but not Pfu. In monocyte-derived macrophages (MDM) infected with VSV-G–pseudotyped HIV-GFP, Pfu amplified zero of two 1,075-bp regions, one of seven 75-bp regions, and three of eight 25-bp regions (Fig. 2*B*). HIV DNA contains 133 75-bp regions, of which 6 of 7 (86%) were Pfu PCR negative and are expected to contain at least two uracils. This suggests that there are at least 133 × 6/7 × 2 or 228 uracils per 10-kb HIV DNA. Performing a similar calculation on the basis of the 25-bp amplicons of which five of eight were Pfu PCR negative, we estimate that there are at least 400 × 5/8 × 2 or 500 uracils per 10-kb HIV DNA. Smaller fragment size and more sampling should provide better resolution and a more accurate estimate. Pfu PCR negative fragments are not more AT rich than the other fragments in our panel. We repeated this uracil mapping experiment using CD4 T cells infected with VSV-G–pseudotyped HIV-GFP or wild-type HIV_{IIIB} (Fig. 2 *C* and *D*). Analysis of the 75-bp regions suggested a lower limit of 228 and 152 uracils, respectively, and analysis of the 25-bp regions suggested ≥400 uracils per HIV reverse transcript for both viruses. Thus, HIV reverse transcripts are heavily uracilated to a similar extent during infection of both macrophages and T cells.

Because our assay uses DNA extracted from multiple infected cells rather than a single infected cell and the lack of a Pfu signal requires that all of the HIV DNA copies in the sample contain uracil in at least one position on each strand, it is likely that the actual number of uracils in each HIV DNA may be higher than our lower limit estimate of 15–50 per kb. Because HIV DNA is AT rich (~60%), at least 15–50 of 600 AT residues or ~2.5–8.3% of thymidines were replaced by uracils during reverse transcription. This number should be compared with estimates of the ratio of dUTP/dTTP in macrophages and lymphocytes of 20–25% (14), which could be considered an upper limit of the uracil-to-thymidine ratio of HIV DNA synthesized in those cells. In contrast, in human chromosomal DNA it is estimated that 70–200 uracils are generated by cytosine deamination per day (<0.0001 per kb) (28), and $Ung^{-/-}$ mouse embryo fibroblasts have ~2,000 uracils per genome (0.001 per kb) (29). In fact Taq and Pfu similarly amplified HIV chromosomal DNA in infected T cells. Therefore, HIV cytosolic DNA is heavily uracilated in human immune cells relative to chromosomal DNA.

HIV Reverse Transcripts Generated in Uracil-Rich Human Macrophages Are Less Susceptible to Autointegration than Transcripts Generated in Uracil-Poor HeLa Cells. We next examined how uracil incorporation affects the early steps of the HIV life cycle. We first compared the relative numbers of HIV late reverse transcripts, autointegrants, and integrated DNA copies in human MDM (which have dUTP/dTTP ~1/4) (14) to HeLa–CD4 cells (dUTP/dTTP ~1/100) (15) after infection with an equal amount of VSV-G–pseudotyped HIV. Reverse transcription, autointegration, and chromosomal integration were measured 10 and 24 h postinfection (hpi) (Fig. 3*A–C*). HIV infection in both cells generated similar amounts of late RT products, suggesting that HIV reverse transcription is not affected by high cellular uracil content in MDMs. In contrast, autointegration was severely reduced in MDMs (11–13% of that in HeLa–CD4 cells), and chromosomal integration increased in MDMs by three- to four-fold. Taq/Pfu PCR analysis (of a 400-bp region) of HIV cytosolic DNA from infected cells confirmed that HIV DNA is more uracilated in MDMs (Fig. 3*D*). Thus, the high cellular uracil content in MDMs does not appear to impede HIV replication. On the contrary, the uracil environment may be beneficial for HIV to complete the early stages of the life cycle, by inhibiting suicidal autointegration and promoting productive chromosomal integration.

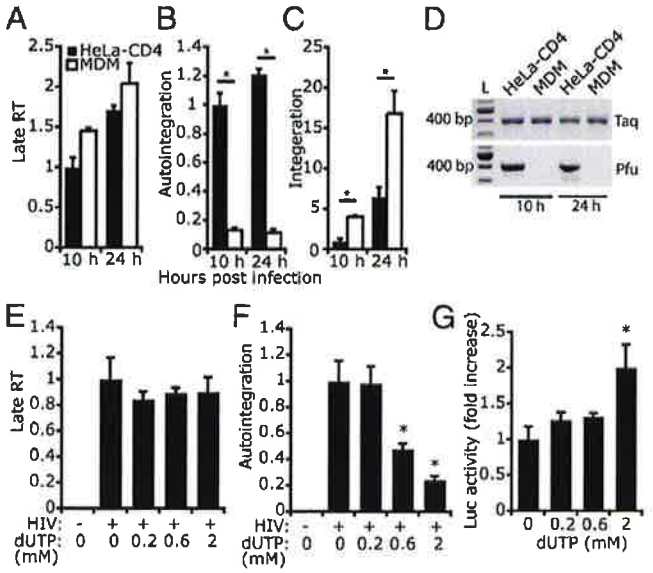


Fig. 3. Uracilation of HIV DNA inhibits autointegration. (*A–D*) HIV RT products produced in infected macrophages are less susceptible to autointegration than those produced in HeLa–CD4 cells. HeLa–CD4 and monocyte-drived macrophages (MDM) were infected with an equal amount of HIV-GFP. Late RT (*A*), autointegration (*B*), and integrated DNA (*C*) were measured 10 h and 24 h postinfection. Late RT was normalized to human mitochondrial DNA; autointegration was normalized to late RT; integrated DNA was normalized to human β-globin DNA. *$P < 0.05$, Student's *t* test. Error bars indicate SD of three independent experiments. Differential uracil incorporation in HIV DNA was verified by Taq/Pfu PCR (*D*). (*E–G*) Culturing HIV-infected HeLa–CD4 cells with increasing amounts of dUTP reduces autointegration and promotes infection. dUTP was added to the medium the night before infection with VSV-G–pseudotyped HIV-Luc virus at a multiplicity of infection of 1. HIV reverse transcription (late RT, *E*) and autointegration (*F*) were measured 10 h postinfection, and Luciferase activity was measured 24 h postinfection (*G*). Values were normalized as above. *$P < 0.05$, Student's *t* test. Error bars indicate SD of three independent experiments.

Uracil Incorporation in HIV DNA Inhibits Autointegration. To further investigate whether uracil incorporation promotes early steps of the HIV life cycle, we added increasing amounts of dUTP to the culture medium of HeLa–CD4 cells and then infected them 1 d later with VSV-G–pseudotyped HIV-Luc and measured HIV reverse transcription and autointegration 10 hpi and Luc activity 24 hpi (Fig. 3 *E–G*). Addition of dUTP did not affect cell growth or viability during the 2-d period of the assay. dUTP treatment also did not affect HIV DNA synthesis in HeLa–CD4 cells (Fig. 3*E*), consistent with the fact that HIV RT does not distinguish between dUTP and dTTP. In contrast, HIV autointegration was inhibited in a dose-dependent manner by adding dUTP (Fig. 3*F*). Luc activity increased twofold in cells treated with the highest dose of dUTP, suggesting that dUTP treatment increased HIV replication, perhaps by reducing suicidal autointegration.

C-to-U Hypermutation by APOBEC3G Inhibits HIV Autointegration. We next examined the effect on HIV autointegration of introducing uracils into HIV DNA by an alternative method, cytosine deamination by A3G (30). A3G is an intrinsic antiviral host protein counteracted by HIV Vif (16). We generated HIV-LucΔvif viruses containing no A3G or increasing amounts of wild-type A3G or A3G-E259Q (an enzymatically defective mutant) (31) by transfecting virus-producing 293T cells with increasing amounts of A3G expression plasmids (Fig. 4*A*). Viral stocks were normalized on the basis of HIV capsid (p24) level measured by ELISA and were used at equal capsid protein levels to infect HeLa–CD4 cells (which do not express A3G). Incorporation of A3G and A3G-E259Q viruses reduced HIV late RT, consistent with a previous report that A3G blocks HIV reverse transcription independently of its deaminase

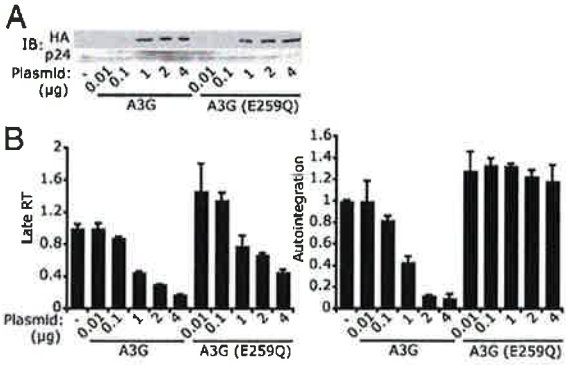


Fig. 4. APOBEC3G-mediated cytosine deamination inhibits HIV autointegration. HeLa–CD4 cells were infected with VSV-G–pseudotyped HIV-LucΔvif viruses containing increasing amounts of wild-type A3G or deaminase-defective A3G (E259Q). The amount of A3G or A3G (E259Q) plasmid used for producing viruses in 293T cells is indicated on the *Bottom*. (*A*) Immunoblot (IB) analysis showing the amount of A3G and A3G (E259Q) in viruses. Both A3G and A3G (E259Q) are HA-tagged. p24 (HIV capsid) was used as a loading control. (*B*) HIV late RT and autointegration were measured 10 h post-infection and normalized as in Fig. 3.

activity (Fig. 4*B* and ref. 32). To assess the effect of increased uracilation of HIV reverse transcripts, we normalized the number of autointegration events to the number of reverse transcripts generated under each condition. Incorporation of enzymatically dead A3G into virions did not affect the amount of autointegration per reverse transcript, whereas wild-type A3G inhibited autointegration. Thus, uracilation of HIV DNA by A3G also inhibited HIV autointegration. However, we did not observe a corresponding increase in overall HIV-Luc activity, perhaps because A3G inhibits multiple steps of the HIV life cycle (32, 33), offsetting its effect of preventing autointegration.

Uracil Incorporation Alters HIV DNA Structure. To understand why HIV DNA uracilation might inhibit autointegration, we first asked whether replacing thymidines with uracils (i.e., removing the 5-methyl group of thymidine) distorts HIV DNA structure. The 5-methyl group of thymidine protrudes into the major groove of DNA and is important for base stacking and structural stability of the DNA. Removing the 5-methyl group increases the minor groove size and affects DNA–protein interactions (34–36). The structure of HIV *gag* DNA (50-bp sequence) synthesized with either thymidines (T-DNA) or uracils in place of thymidine (U-DNA) was compared by circular dichroism (CD) spectrometry and thermal denaturation (Fig. 5). U-DNA displayed a significant blueshift in the 245- to 300-nm range of the CD spectrum compared with T-DNA (Fig. 5*A*). A similar blueshift was also observed when Bromo-dU was used to replace thymidine and has been attributed to turn structure distortion (37). U-DNA also showed reduced thermostability compared with T-DNA by melt curve analysis, in which the release of the SYBR green DNA dye was measured as a function of temperature (Fig. 5*B*, *Upper*). U-DNA consistently displayed a melting temperature 2° to 3° lower than T-DNA (Fig. 5*B*, *Lower*), indicating that uracil-containing DNA has reduced thermostability. Taken together, we conclude that uracilation significantly alters the structure and stability of DNA.

Uracil-DNA Inhibits Integration by HIV IN. We next examined whether uracil-induced DNA structural changes in target DNA might inhibit HIV IN strand transfer. We used an in vitro integration assay in which recombinant HIV IN catalyzes half-site integration (joining of only one strand of dsDNA) of a donor and a target DNA (38). This assay is often used to characterize IN strand transfer activity and assess host factors, such as LEDGF, that influence integration (39). In the assay, the donor DNA (50 bp, HIV U5 sequence) contains a recessed CA dinucleotide end that mimics the processed viral DNA end. The target DNA

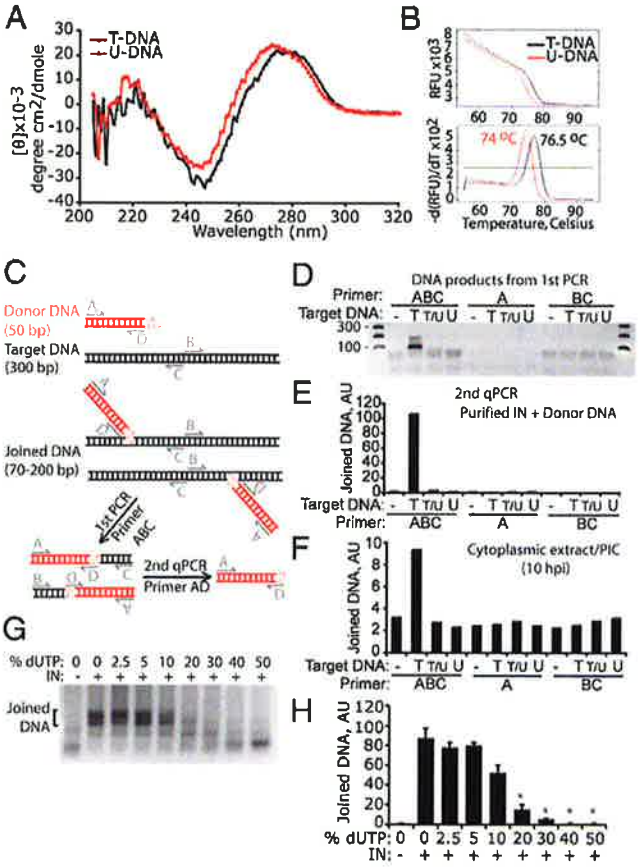


Fig. 5. Uracil inhibits IN-mediated integration. (*A* and *B*) Uracil distorts HIV DNA as assessed by circular dichroism (*A*) and melt curve (*B*) analysis of a synthetic 50 bp (HIV *gag* sequence, 48% GC) dsDNA, containing thymidines (T-DNA) or uracils in place of thymidines (U-DNA). See *Materials and Methods* for sequences. Melt curve plot was generated by measuring the dissociation of SYBR green dye from dsDNA as a function of temperature (*B*, *Upper*). A derivative plot (*B*, *Lower*) is used to calculate the melting temperature (labels on peaks). Results in *A* and *B* are representative of three experiments. (*C–F*) Target DNA uracilation inhibits IN-mediated integration. (*C*) Diagram of half-site integration assay catalyzed by recombinant HIV IN. The donor DNA (preprocessed to have the 5′-AC dinucleotide overhang) is in red and the target DNA is in black. (*D*) DNA gel electrophoresis of PCR amplified integration products. Target DNA used in the in vitro integration assay contained either all thymidine (T), an equal mixture of thymidine and uracil (T/U), or all uracil (U). Primers used for amplification are shown on *Top*. Expected donor-target joining product length ranges from 70 to 200 bp. (*E*) qPCR analysis (with primers A and D) of PCR products obtained in *D*. AU, arbitrary unit (same below). (*F*) In vitro integration assay was carried out as in *E* using HIV PICs isolated from infected HeLa–CD4 cells to replace recombinant IN and donor DNA. Results in *D–F* are representative of two experiments. (*G* and *H*) Ten to 20% of uracil in target DNA inhibits integration. Dose–response experiment using varying levels of uracil in target DNA (as indicated) in an in vitro integration assay as in *D*. DNA gel electrophoresis of PCR-amplified integration products is shown in *G*, and quantification of the joined DNA, analyzed by qPCR as in *E* for two independent experiments, is shown in *H*. % dUTP, percentage of dUTP of total dUTP plus dTTP. Data represent mean ± SD. $^*P < 0.01$, Student's *t* test (compared with 0% dUTP plus IN).

contains non-HIV sequence (300 bp, AT = 53%; *Materials and Methods*), which allowed us to perform the assay with either recombinant IN and synthesized donor DNA or with HIV PIC isolated from infected cells. Target DNA samples that contain all T, half T/half U, or all U were used to determine the effect on integration of uracil in the target strand. U5 DNA integration products were detected by a semiquantitative nested PCR assay [similar to the *Alu*-PCR method used for quantifying integrated HIV DNA (ref. 40 and Fig. 5*C*)]. In the first PCR, a successfully

MICROBIOLOGY

integrated product containing both a donor and a target DNA molecule can be amplified only by using all three primers (ABC), but not by A or BC alone. PCR products from the first PCR step are heterogeneous in length (Fig. 5*D*), representing integration events that occurred at multiple sites along the target DNA. All DNA products from the first PCR contain the donor DNA sequence, which was then amplified by "nested" qPCR (with primers A and D). Recombinant HIV IN robustly catalyzed half-site integration into target DNA that contained no uracil, as previously described (Fig. 5 *D* and *E* and refs. 38, 39). In contrast, half T and half U (T/U) or all U in the target DNA completely inhibited integration, suggesting that IN strongly disfavors uracil-containing target DNA. To confirm these results, cytoplasmic extract from infected HeLa–CD4 cells (10 hpi) was used as a source of native HIV IN and donor DNA, which was incubated with the same target DNA oligonucleotides as above, containing varying T/U ratios. Although PICs support efficient concerted integration of endogenous U3 and U5 donor DNA ends (41), our assay monitored U5 end joining only. Integration occurred with T-DNA, but not with T/U- or U-DNA (Fig. 5*F*). Although HIV PIC IN formed only a fraction of integration products compared with the recombinant enzyme, these results clearly demonstrate that uracilation of target DNA inhibits HIV integration in vitro.

To determine how much uracil in target DNA is required to inhibit integration, we performed a dose–response in vitro integration experiment using varying levels of dU in the target DNA (Fig. 5 *G* and *H*). Integration was inhibited by only 10% dU (of total dT + dU) in target DNA; 20% dU almost completely blocked integration. Because at least 8%, and possibly as much as 25%, of thymidines in HIV DNA in macrophages and CD4 T cells are replaced by uracil, our data suggest that HIV uracilation contributes to blocking autointegration in a physiologically relevant way.

Discussion

In this study, we provide direct evidence that HIV DNA generated in infected primary human immune cells is heavily uracilated and that uracilation protects HIV DNA from the suicidal pathway of autointegration, facilitating chromosomal integration. Our results suggest that at least 8% of thymidines in HIV DNA in macrophages and CD4 T cells are replaced by uracils. A more accurate estimate of the degree of uracilation could be obtained by refining the Taq/Pfu PCR assay to amplify smaller segments than the smallest region we examined (25 bp) using fewer infected cells. In the immune cells that HIV infects, uracil: thymidine ratios approach 25%, and because HIV RT uses either base equivalently, freshly reverse transcribed HIV DNA likely contains about 20–25% uracils in place of thymidine. Several studies have suggested that UNG2, the major mammalian UNG, is packaged into HIV virions. However, it is unknown whether UNG2 is contained within the reverse transcription complex and retained as it matures into the PIC or whether it acts on nascent HIV reverse transcripts. Our results do not exclude the possibility that viral-associated UNG2 is active. However, much of the incorporated uracil in HIV DNA is retained in cytosolic HIV DNA. Virions produced in cells lacking UNG2 activity have no replication defect (24), indicating that HIV tolerates a high degree of uracilation. Because BER is highly efficient within the nucleus, it is likely that incorporated uracils in integrated chromosomal HIV DNA are efficiently removed and faithfully replaced by thymidines once integration has occurred.

HIV IN favors integration (for both integration and autointegration) (10, 42, 43) into target DNAs at sites that loosely match an 11-bp consensus sequence containing approximately four to seven thymidines. Thus, many available autointegration sites would contain at least one nearby uracil. We showed here that uracilation of target DNA inhibits IN strand transfer. Thus, uracil-rich HIV cytosolic DNA is a poor target for (auto)integration compared with uracil-poor chromosomal DNA. Lentiviral IN may have evolved to enhance chromosomal integration in uracil-rich cells. Structural changes caused by uracilation could have a detrimental effect on IN targeting. Indeed, the recently solved structure of the prototype foamy virus (PFV) IN multimer in complex with DNA oligomers revealed that the major groove is severely widened and minor groove compressed at the site of integration (44, 45). Replacing thymidines with uracils, which involves removal of the 5-methyl group of thymidines, expands the minor groove, thereby potentially creating structural hindrance for IN access to inhibit strand transfer. 5-methyl groups of thymidines lying within the core of the target DNA integration site directly contact PFV IN in the cocrystal structure (44). Thus, replacing these thymidines with uracils could affect IN binding.

In addition to a role in protecting HIV from autointegration, uracilation of HIV reverse transcripts might influence other events in the early phase of HIV replication. We recently found that HIV DNA can be recognized by an unknown innate immune DNA cytosolic sensor in macrophages and T cells to trigger production of type I interferons. Cytosolic HIV DNA uracilation might interfere with nucleic acid detection in the cytosol that could potentially activate innate immunity (46).

We previously showed that the host SET complex inhibits HIV autointegration by binding to HIV DNA soon after reverse transcription in the cytosol (10). Here we found that intrinsic uracilation of HIV DNA in natural target cells also reduces autointegration. Both processes (SET complex binding and HIV DNA uracilation) can independently inhibit autointegration. For example, the SET complex inhibits HIV autointegration in HeLa–CD4 cells, where HIV DNA is not uracilated (10), and uracilated target DNA blocks IN-mediated strand transfer in vitro in the absence of the SET complex (Fig. 5*D*). However, it is likely that these two mechanisms also act in concert. The SET complex has BER activity (11, 12). Because one BER function is to repair misincorporated uracils, the SET complex may recognize and bind HIV DNA via its uracils. In fact, HMGB2, a component of the SET complex, preferentially binds to distorted DNA structures. We envision a model in which HIV DNA is uracilated during reverse transcription, which attracts the SET complex to bind. SET complex binding to these uracilated regions of HIV DNA in the cytosol then shields them from autointegration. It may be that the SET complex does not repair HIV DNA uracils in the cytosol, where the abundance of other downstream BER pathway enzymes may be limiting. Because the SET complex shuttles between the cytosol and the nucleus (47), it may traffic with HIV DNA into the nucleus for integration and then mark the spots of uracil incorporation for BER within the nucleus. Further studies are needed to investigate these possibilities.

Materials and Methods

Human $CD4^+$ T cells and macrophages were isolated from PBMC as in ref. 46. HIV_{IIIB}, HIV-Luc, and HIV-GFP were generated as in ref. 46. HIV DNA products were analyzed as in ref. 10. In vitro integration assay was performed as in ref. 38. Taq/Pfu PCR details and other additional information are available in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Peter Cherepanov (Imperial College London) for insightful discussions, Dana Gabuzda (Dana-Farber Cancer Institute) for HIV-GFP plasmid, Ned Landau (New York University) for A3G plasmid, Stephen Blacklow (Harvard Medical School) for help with the CD measurement, and members of the J.L. laboratory for helpful discussions. This work was supported by National Institutes of Health Grants AI45587 (to J.L.), T32 HL066987-10 (to N.Y.), and AI052014 (to A.E.) and an Endowed Scholars Award from University of Texas Southwestern (to N.Y.).

1. Li Y, et al. (1991) Molecular characterization of human immunodeficiency virus type 1 cloned directly from uncultured human brain tissue: Identification of replication-competent and -defective viral genomes. *J Virol* 65:3973–3985.
2. Lee MS, Craigie R (1998) A previously unidentified host protein protects retroviral DNA from autointegration. *Proc Natl Acad Sci USA* 95:1528–1533.
3. Suzuki Y, Craigie R (2002) Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes. *J Virol* 76:12376–12380.
4. Garfinkel DJ, et al. (2006) Retrotransposon suicide: Formation of Ty1 circles and autointegration via a central DNA flap. *J Virol* 80:11920–11934.

5. Benjamin HW, Kleckner N (1989) Intramolecular transposition by Tn10. *Cell* 59: 373–383.
6. Maxwell A, Craigie R, Mizuuchi K (1987) B protein of bacteriophage mu is an ATPase that preferentially stimulates intermolecular DNA strand transfer. *Proc Natl Acad Sci USA* 84:699–703.
7. Lee MS, Craigie R (1994) Protection of retroviral DNA from autointegration: Involvement of a cellular factor. *Proc Natl Acad Sci USA* 91:9823–9827.
8. Chen H, Engelman A (1998) The barrier-to-autointegration protein is a host factor for HIV type 1 integration. *Proc Natl Acad Sci USA* 95:15270–15274.
9. Skoko D, et al. (2009) Barrier-to-autointegration factor (BAF) condenses DNA by looping. *Proc Natl Acad Sci USA* 106:16610–16615.
10. Yan N, Cherepanov P, Daigle JE, Engelman A, Lieberman J (2009) The SET complex acts as a barrier to autointegration of HIV-1. *PLoS Pathog* 5:e1000327.
11. Postel EH, Abramczyk BM, Levit MN, Kyin S (2000) Catalysis of DNA cleavage and nucleoside triphosphate synthesis by NM23-H2/NDP kinase share an active site that implies a DNA repair function. *Proc Natl Acad Sci USA* 97:14194–14199.
12. Höss M, et al. (1999) A human DNA editing enzyme homologous to the Escherichia coli DnaQ/MutD protein. *EMBO J* 18:3868–3875.
13. Traut TW (1994) Physiological concentrations of purines and pyrimidines. *Mol Cell Biochem* 140:1–22.
14. Aquaro S, et al. (2002) Macrophages and HIV infection: Therapeutical approaches toward this strategic virus reservoir. *Antiviral Res* 55:209–225.
15. Mahagaokar S, Orengo A, Rao PN (1980) The turnover of deoxyuridine triphosphate during the HeLa cell cycle. *Exp Cell Res* 125:86–94.
16. Chiu YL, Greene WC (2008) The APOBEC3 cytidine deaminases: An innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol* 26:317–353.
17. Pyles RB, Thompson RL (1994) Evidence that the herpes simplex virus type 1 uracil DNA glycosylase is required for efficient viral replication and latency in the murine nervous system. *J Virol* 68:4963–4972.
18. Ellison KS, Peng W, McFadden G (1996) Mutations in active-site residues of the uracil-DNA glycosylase encoded by vaccinia virus are incompatible with virus viability. *J Virol* 70:7965–7973.
19. Millns AK, Carpenter MS, DeLange AM (1994) The vaccinia virus-encoded uracil DNA glycosylase has an essential role in viral DNA replication. *Virology* 198:504–513.
20. Lichtenstein DL, et al. (1995) Replication in vitro and in vivo of an equine infectious anemia virus mutant deficient in dUTPase activity. *J Virol* 69:2881–2888.
21. Wagaman PC, et al. (1993) Molecular cloning and characterization of deoxyuridine triphosphatase from feline immunodeficiency virus (FIV). *Virology* 196:451–457.
22. Selig L, et al. (1997) Uracil DNA glycosylase specifically interacts with Vpr of both human immunodeficiency virus type 1 and simian immunodeficiency virus of sooty mangabeys, but binding does not correlate with cell cycle arrest. *J Virol* 71:4842–4846.
23. Willetts KE, et al. (1999) DNA repair enzyme uracil DNA glycosylase is specifically incorporated into human immunodeficiency virus type 1 viral particles through a Vpr-independent mechanism. *J Virol* 73:1682–1688.
24. Kaiser SM, Emerman M (2006) Uracil DNA glycosylase is dispensable for human immunodeficiency virus type 1 replication and does not contribute to the antiviral effects of the cytidine deaminase Apobec3G. *J Virol* 80:875–882.
25. Lasken RS, Schuster DM, Rashtchian A (1996) Archaebacterial DNA polymerases tightly bind uracil-containing DNA. *J Biol Chem* 271:17692–17696.
26. Hogrefe HH, Hansen CJ, Scott BR, Nielson KB (2002) Archaeal dUTPase enhances PCR amplifications with archaeal DNA polymerases by preventing dUTP incorporation. *Proc Natl Acad Sci USA* 99:596–601.
27. Horváth A, Vértessy BG (2010) A one-step method for quantitative determination of uracil in DNA by real-time PCR. *Nucleic Acids Res* 38:e196.
28. Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362: 709–715.
29. Nilsen H, et al. (2000) Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. *Mol Cell* 5:1059–1065.
30. Harris RS, et al. (2003) DNA deamination mediates innate immunity to retroviral infection. *Cell* 113:803–809.
31. Newman EN, et al. (2005) Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity. *Curr Biol* 15:166–170.
32. Bishop KN, Verma M, Kim E-Y, Wolinsky SM, Malim MH (2008) APOBEC3G inhibits elongation of HIV-1 reverse transcripts. *PLoS Pathog* 4:e1000231.
33. Mbisa JL, et al. (2007) Human immunodeficiency virus type 1 cDNAs produced in the presence of APOBEC3G exhibit defects in plus-strand DNA transfer and integration. *J Virol* 81:7099–7110.
34. Tate PH, Bird AP (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* 3:226–231.
35. Bailly C, Crow S, Minnock A, Waring MJ (1999) Demethylation of thymine residues affects DNA cleavage by endonucleases but not sequence recognition by drugs. *J Mol Biol* 291:561–573.
36. Marathe A, Bansal M (2010) The 5-methyl group in thymine dynamically influences the structure of A-tracts in DNA at the local and global level. *J Phys Chem B* 114: 5534–5546.
37. Augenlicht L, Nicolini C, Baserga R (1974) Circular dichroism and thermal denaturation studies of chromatin and DNA from BrdU-treated mouse fibroblasts. *Biochem Biophys Res Commun* 59:920–926.
38. Dar MJ, et al. (2009) Biochemical and virological analysis of the 18-residue C-terminal tail of HIV-1 integrase. *Retrovirology* 6:94.
39. Hare S, et al. (2009) Structural basis for functional tetramerization of lentiviral integrase. *PLoS Pathog* 5:e1000515.
40. O'Doherty U, Swiggard WJ, Jeyakumar D, McGain D, Malim MH (2002) A sensitive, quantitative assay for human immunodeficiency virus type 1 integration. *J Virol* 76: 10942–10950.
41. Chen H, Engelman A (2001) Asymmetric processing of human immunodeficiency virus type 1 cDNA in vivo: Implications for functional end coupling during the chemical steps of DNA transposition. *Mol Cell Biol* 21:6758–6767.
42. Holman AG, Coffin JM (2005) Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci USA* 102:6103–6107.
43. Shun MC, et al. (2007) LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev* 21:1767–1778.
44. Maertens GN, Hare S, Cherepanov P (2010) The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 468:326–329.
45. Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P (2010) Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* 464:232–236.
46. Yan N, Regalado-Magdos AD, Stiggelbout B, Lee-Kirsch MA, Lieberman J (2010) The cytosolic exonuclease TREX1 inhibits the innate immune response to human immunodeficiency virus type 1. *Nat Immunol* 11:1005–1013.
47. Chowdhury D, Lieberman J (2008) Death by a thousand cuts: Granzyme pathways of programmed cell death. *Annu Rev Immunol* 26:389–420.

OPEN ACCESS Freely available online

PLoS GENETICS

Conserved Regulation of p53 Network Dosage by MicroRNA–125b Occurs through Evolving miRNA–Target Gene Pairs

Minh T. N. Le[1,2,3☯], Ng Shyh-Chang[1,4☯], Swea Ling Khaw[1,5], Lingzi Chin[1], Cathleen Teh[6], Junliang Tay[1], Elizabeth O'Day[2], Vladimir Korzh[5], Henry Yang[7], Ashish Lal[2,8], Judy Lieberman[2], Harvey F. Lodish[3,9,10]*, Bing Lim[1,3,11]*

1 Stem Cell and Developmental Biology, Genome Institute of Singapore, Singapore, Singapore, **2** Immune Disease Institute and Program in Cellular and Molecular Medicine, Children's Hospital Boston, Harvard Medical School, Boston, Massachusetts, United States of America, **3** Computation and Systems Biology, Singapore–MIT Alliance, Singapore, Singapore, **4** Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America, **5** NUS Graduate School for Integrative Sciences and Engineering, Singapore, Singapore, **6** Fish Developmental Biology, Institute of Molecular and Cell Biology, Singapore, Singapore, **7** Bioinformatics Group, Singapore Immunology Network, Singapore, Singapore, **8** Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **9** Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, United States of America, **10** Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **11** Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

MicroRNAs regulate networks of genes to orchestrate cellular functions. *MiR-125b*, the vertebrate homologue of the *Caenorhabditis elegans* microRNA *lin-4*, has been implicated in the regulation of neural and hematopoietic stem cell homeostasis, analogous to how *lin-4* regulates stem cells in *C. elegans*. Depending on the cell context, *miR-125b* has been proposed to regulate both apoptosis and proliferation. Because the p53 network is a central regulator of both apoptosis and proliferation, the dual roles of *miR-125b* raise the question of what genes in the p53 network might be regulated by *miR-125b*. By using a gain- and loss-of-function screen for *miR-125b* targets in humans, mice, and zebrafish and by validating these targets with the luciferase assay and a novel miRNA pull-down assay, we demonstrate that *miR-125b* directly represses 20 novel targets in the p53 network. These targets include both apoptosis regulators like *Bak1*, *Igfbp3*, *Itch*, *Puma*, *Prkra*, *Tp53inp1*, *Tp53*, *Zac1*, and also cell-cycle regulators like *cyclin C*, *Cdc25c*, *Cdkn2c*, *Edn1*, *Ppp1ca*, *Sel1l*, in the p53 network. We found that, although each miRNA–target pair was seldom conserved, miR-125b regulation of the p53 pathway is conserved at the network level. Our results lead us to propose that *miR-125b* buffers and fine-tunes p53 network activity by regulating the dose of both proliferative and apoptotic regulators, with implications for tissue stem cell homeostasis and oncogenesis.

Citation: Le MTN, Shyh-Chang N, Khaw SL, Chin L, Teh C, et al. (2011) Conserved Regulation of p53 Network Dosage by MicroRNA–125b Occurs through Evolving miRNA–Target Gene Pairs. PLoS Genet 7(9): e1002242. doi:10.1371/journal.pgen.1002242

Editor: Michael T. McManus, University of California San Francisco, United States of America

Received March 6, 2011; **Accepted** June 27, 2011; **Published** September 15, 2011

Funding: MTNL, NS-C, SLK, LC, JT, CT, VK, and BL were supported by ASTAR, Singapore. MTNL was a L'Oréal Singapore for Women in Science National Fellow. BL and HFL were partially supported by SMA grant C-382-641-001-091. HFL was supported by NIH grant R01 DK068348. JL, EO, MTNL, and AL were supported by NIH. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: limb1@gis.a-star.edu.sg, blim@bidmc.havard.edu (BL); lodish@wi.mit.edu (HFL)

☯ These authors contributed equally to this work.

Introduction

MicroRNAs (miRNAs) are short non-coding RNA molecules that were first discovered as regulators of developmental timing, and later found to regulate complex networks of genes to orchestrate cellular functions. *Lin-4* was the first miRNA gene to be discovered, and shown to regulate developmental timing by repressing its target genes at the post-transcriptional level [1]. Subsequently, miRNAs were found to regulate processes ranging from proliferation and apoptosis, to cell differentiation and signal transduction [2–4]. Several miRNAs are conserved in metazoan evolution, one prominent example being *lin-4* whose vertebrate homologues comprise the *miR-125a/b* family [5]. Much like *lin-4*'s role of regulating the homeostasis of reiterative or self-renewing stem cells in *C. elegans* [6], recent studies have shown that *miR-125a/b* regulates mammalian neural stem cell commitment, as well as the mammalian hematopoietic stem cell (HSC) pool size [7–10]. Although *Lin28* and *Bak1* have been proposed as the critical targets of *miR-125a/b* for regulating these stem cell compartments [8,9], the hundreds of predicted targets for *miR-125a/b* suggest a more complex interplay between *miR-125a/b* and its targets in regulating proliferation and differentiation.

Depending on the cell context, *miR-125b* has been proposed to regulate both apoptosis and proliferation. *miR-125b* has been shown to downregulate apoptosis in many contexts, in some cases by repressing Tp53 and Bak1. Examples include mammalian hematopoietic stem cells, human leukemia cells, neuroblastoma cells, breast cancer and prostate cancer cells [9–18]. During zebrafish embryogenesis, loss of *miR-125b* leads to widespread apoptosis in a p53-dependent manner, causing severe defects in neurogenesis and somitogenesis [16]. On the other hand, *miR-*

Author Summary

MicroRNAs (miRNAs) are tiny endogenous RNAs that can regulate the expression of hundreds of genes simultaneously, thus orchestrating changes in gene networks and mediating cellular functions in both plants and animals. Although the identification of individual targets of miRNAs is of major importance, to date few studies have sought to uncover miRNA targets at the gene network level and general principles of miRNA regulation at the network level. Here we describe how *miR-125b* targets 20 apoptosis and proliferation genes in the p53 network. We found that, although each miRNA-target pair evolves rapidly across vertebrates, regulation of the p53 pathway by *miR-125b* is conserved at the network level. The structure of the *miR-125b* regulatory network suggests that *miR-125b* buffers and fine-tunes p53 network activity. This buffering feature of *miR-125b* has implications for our understanding of how *miR-125b* regulates oncogenesis and tissue stem cell homeostasis. We believe these findings on *miR-125b* support a new fundamental principle for how miRNAs regulate gene networks in general.

125b can also downregulate proliferation in a variety of human cancer cell-lines [19–23] and one of its bona fide targets Lin28, also promotes cancer cell proliferation [24]. Therefore in different contexts, *miR-125b* appears to be able to regulate both apoptosis and proliferation.

Another molecular pathway that regulates both apoptosis and proliferation is the highly conserved p53 network [25–28]. Due to the central role of the p53 network in these two processes, and because we found that *miR-125b* regulates both human and zebrafish Tp53 but not mouse Tp53 [16], we sought to examine if *miR-125b* regulates the p53 network in a conserved manner in vertebrates. To address this question, we used a gain- and loss-of-function screen for *miR-125b* targets in different vertebrates, and validated these targets with the luciferase assay and a novel miRNA-target pull-down assay. We demonstrate that *miR-125b* directly represses 20 novel targets in the p53 network, including both apoptosis regulators like *Bak1*, *Igfbp3*, *Itch*, *Puma*, *Prkra*, *Tp53inp1*, *Tp53*, *Zac1*, and also cell-cycle regulators like *cyclin C*, *Cdc25c*, *Cdkn2c*, *Edn1*, *Ppp1ca*, *Sell1*. We found that although individual miRNA-target pairs were seldom conserved, regulation of the p53 network by *miR-125b* appears to be conserved at the network-level. This led us to propose that *miR-125b* buffers and fine-tunes p53 network dosage, with implications for the role of *miR-125b* in tissue stem cell homeostasis and oncogenesis.

Results

Identifying direct targets of miR-125b in the p53 network

To systematically identify direct targets of *miR-125b* in the p53 network of vertebrates, we first employed a bioinformatics approach by identifying all predicted *miR-125b* targets in the p53 network, followed by three complementary methods to screen and validate these targets for both direct binding and repression by *miR-125b* (Figure 1). Existing databases and prediction algorithms were used to shortlist a set of p53 network genes predicted to possess *miR-125b*-binding sites in their 3′ UTRs. We analyzed the Ingenuity Pathways Analysis™ (IPA) database and the p53 Knowledgebase [29,30] for a list of genes and proteins that participate in the p53 network, either by regulating p53 upstream, by direct interaction with p53 protein, or by serving as effectors of p53 function downstream. We then analyzed the TargetScan and

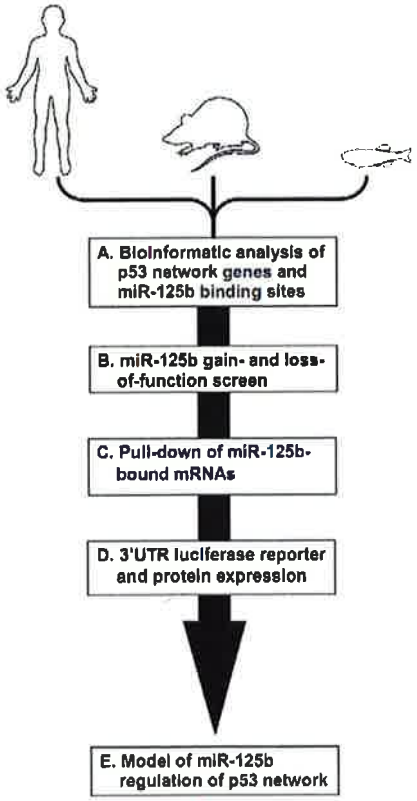


Figure 1. Identifying miR-125b targets in the p53 network of vertebrates. Schematic of experimental design and workflow. (A) Bioinformatic analysis was performed on p53 network genes listed in the Ingenuity Pathways Analysis database and p53 Knowledgebase, and miR-125b binding sites predicted by the TargetScan and MicroCosm databases. (B) p53 network genes were screened for miR-125b targets by using gain- (GOF) and loss-of-function (LOF) of miR-125b in human cells, mouse cells and zebrafish embryos, as indicated by effects on gene expression using qRT-PCR. (C) p53 network genes that were positive in either the GOF or LOF screen were assayed for direct binding to miR-125b using a biotinylated microRNA pull-down method. (D) p53 network genes that were also positive in the miR-125b pull-down were finally validated as miR-125b targets by 3′ UTR luciferase reporter assays and Western blots for protein expression. (E) A model of how miR-125b regulates the p53 network across vertebrates was constructed using our combined datasets for human, mouse and zebrafish cells.
doi:10.1371/journal.pgen.1002242.g001

MicroCosm Target databases [31,32] for genes that are predicted to possess *miR-125b*-binding sites in their 3′ UTRs, in three vertebrate genomes: human, mouse and zebrafish. The genes at the intersection of the predicted miR-125b target list and the list of p53 network genes constituted our list of predicted *miR-125b* targets in the p53 network (Table S1).

miR-125b gain- and loss-of-function screen in 3 vertebrates

Next we sought to screen our list of predicted targets for significant repression by *miR-125b* in cells, by performing a *miR-125b* gain- and loss-of-function screen. Gain-of-function (GOF) in *miR-125b* was achieved by transfection of *miR-125b* duplex into human SH-SY5Y or mouse N2A neuroblastoma cells, whereas loss-of-function (LOF) in *miR-125b* was achieved in human primary lung fibroblasts or mouse 3T3 fibroblasts by knocking down *miR-125b* with an antisense (AS) RNA (Figure 2A). We chose to perform a gain-of-function screen in human (SH-SY5Y) or mouse (N2A) neuroblastoma cells, because these cells possess low

levels of endogenous *miR-125b* (Figure S1A, S1B). For the loss-of-function screen, we chose human fetal lung (hLF) or mouse (3T3) fibroblasts because they possess high levels of *miR-125b* (Figure S1C, S1D). *miR-125a*-AS was co-transfected with *miR-125b*-AS to achieve a complete silencing of the *miR-125a/b* family, because *miR-125a*, which shares the same seed sequence and the same predicted targets as *miR-125b*, is also highly expressed in human and mouse fibroblasts (Figure S1C, S1D). Genes that were either significantly repressed by *miR-125b* or significantly derepressed by *miR-125a/b*-AS with fold-changes within the range of microRNA regulation ($P<0.05$, fold change > 1.3), were selected as candidate *miR-125b* targets (Figure 2B–2D). For zebrafish embryos, which possess high levels of *miR-125b*, the loss-of-function (LOF) screen was performed using an antisense morpholino cocktail that blocks the loop regions of all 3 pre-*miR-125b* hairpin precursors [16]. The gain-of-function (GOF) screen was performed by co-injecting *miR-125b* duplex with the morpholino (Figure 2A). All gene expression changes were measured with at least three biological replicates using qRT-PCR.

Our GOF/LOF screen revealed that in humans, out of 29 predicted targets in the p53 network, 13 genes were derepressed by *miR-125a/b*-AS in hLF cells and 20 genes were repressed by *miR-125b* in SH-SY5Y cells (Figure 2B). In mice, out of 22 predicted targets in the p53 network, 11 genes were derepressed by *miR-125a/b*-AS in 3T3 cells and 12 genes were repressed by *miR-125b* in N2A cells (Figure 2C). In zebrafish embryos, out of 20 predicted targets in the p53 network, 13 genes were derepressed by pre-*miR-125b* morpholino and 12 genes were repressed by the injection of *miR-125b* duplex (Figure 2D). In total, 22 human genes, 13 mouse genes and 14 zebrafish genes passed the gain- and loss-of-function qRT-PCR screen.

Direct binding interactions between miR-125b and mRNA targets from the p53 network

To assess which candidate *miR-125b* targets identified in the gain- and loss-of-function qRT-PCR screen are directly bound by *miR-125b* in cells, we employed a novel miRNA pull-down method developed by Lal *et al.* (manuscript in preparation). RNA transcripts bound to biotinylated-miR-125b were pulled down with streptavidin beads and quantified by qRT-PCR relative to mRNAs bound to biotinylated-control miRNA ($\log_2$ fold change > 0.5, $P<0.05$). In this assay, biotinylated miRNAs were shown to be loaded into the RNA-induced silencing complex (RISC) and fully functional in repressing their target mRNAs (Lal *et al.*, manuscript in preparation). This method provides a robust and complementary method for detecting miRNAs bound to endogenous target mRNAs, and serves as a useful approach for distinguishing direct and indirect targets in the same pathway (Lal *et al.*, manuscript in preparation). Quantification of the pulled down mRNA targets in hLF cells revealed that 13 out of 22 gene transcripts, *Bak1, Cdc25c, Edn1, Igfbp3, Mre11a, Ppp1ca, Ppp2ca, Prkra, Puma, Tdg, Tp53, Tp53inp1* and *Zac1*, were direct binding targets of *miR-125b* in human cells (Figure 3A). In mouse 3T3 cells, 11 out of 13 gene transcripts, *Bak1, Hspa5, Itch, Ppp1ca, Ppp2ca, Prkra, Puma, Sel1l, Sp1, Tdg* and *Tp53inp1*, were found to be direct binding targets of *miR-125b* (Figure 3B). In zebrafish embryos, 8 out of 14 gene transcripts, *Cdc25c, Cdkn2c, Gtf2h1, Hspa5, Itch, Ppp1ca, Sel1l,* and *Tp53*, were pulled down by *miR-125b* (Figure 3C). *Tp53* mRNA was pulled down by *miR-125b* only in human lung fibroblasts and zebrafish embryos but not in mouse fibroblasts, consistent with previously published results [16] and the Targetscan algorithmic prediction that *miR-125b* targets *Tp53* in humans and zebrafish but not in mice.

Validation of miR-125b targets in the p53 network

As a final validation of the candidate *miR-125b* targets we have identified in the p53 network, we tested our candidate target genes with the luciferase reporter assay. Where cloning was successful, we cloned the entire 3′ UTR of selected candidate target genes into a *Renilla* luciferase reporter, and assayed luciferase expression following co-transfection of *miR-125b* duplex into HEK-293T cells. Transfection of *miR-125b* significantly suppressed 40-60% ($P<0.01$) of the luciferase activity of many 3′ UTR reporters of the *miR-125b* targets we analyzed, relative to transfection of the negative control miRNA (Figure 4A). For humans, the 3′ UTR reporters of *Bak1, Cdc25c, Ppp1ca, Ppp2ca, Prkra, Puma, Tdg, Tp53, Tp53inp1,* and *Zac1* were significantly suppressed by *miR-125b*. In mice, the 3′ UTR reporters of *Bak1, Itch, Ppp1ca, Ppp2ca, Prkra, Puma, Sel1l, Tdg,* and *Tp53inp1* were significantly suppressed by *miR-125b* (Figure 4B). In zebrafish, the 3′ UTR reporters of *Ccnc, Cdc25c, Cdkn2c, Gtf2h1, Hspa5, Ppp1ca,* and *Tp53* were significantly suppressed by *miR-125b* (Figure 4C). With the exception of zebrafish *Ccnc*, all genes tested were positive in the *miR-125b*-pull-down as well as the *miR-125b* gain- and loss-of-function screen. Amongst these targets, we found *Ppp1ca, Prkra* and *Tp53* to be especially interesting from the evolutionary viewpoint, since all 3 vertebrate species possess these 3 genes, but each gene shows a different pattern of evolutionary conservation with respect to *miR-125b*-repression. *Ppp1ca* is repressed by *miR-125b* in all 3 species, *Prkra* is repressed by *miR-125b* in humans and mice, while *Tp53* is repressed in humans and zebrafish. To examine the sequence evolution of these miRNA-mRNA pairs in greater detail, we compared the Targetscan-predicted *miR-125b* binding sites of these genes in humans, mice and zebrafish. In *Ppp1ca*, the predicted binding site is 95% identical between humans and mice and 55% identical between humans and zebrafish, while the seed binding sequence is 100% conserved in all 3 species (Figure 4D). In *Prkra*, the predicted binding site is 94% identical between humans and mice, but only 26% identical between humans and zebrafish, while the seed binding sequence is completely absent in zebrafish (Figure 4D). In contrast, the predicted binding site in *Tp53* is 64% identical between humans and zebrafish, and the seed binding sequence is 100% conserved between humans and zebrafish, but only 36% identical between humans and mice, while the mouse seed binding sequence has acquired 2 point mutations (Figure 4D). The *miR-125b*-repression patterns we observed for each of these genes in the qPCR, pull-down and luciferase assays are consistent with these DNA sequence analyses, suggesting that evolution in the miRNA-mRNA binding is driving the evolution in *miR-125b*-repression patterns. Introduction of point mutations into the predicted seed binding sequences abrogated *miR-125b*-repression of each target 3′UTR luciferase reporter ($P<0.05$), validating the predicted *miR-125b* binding sites and confirming the miRNA-mRNA sequence evolution patterns we observed (Figure 4E).

Finally, we checked *miR-125b* regulation of protein expression in a subset of p53 network targets for which reliable Western blotting was possible. *miR-125b* significantly downregulated the protein levels of human BAK1, PPP1CA, TP53INP1, PPP2CA, CDC25C, and TP53 in SH-SY5Y neuroblastoma cells (Figure 4F). In mouse N2A neuroblastoma cells, *miR-125b* significantly downregulated mouse BAK1, PPP1CA, PUMA, and ITCH protein (Figure 4G).

miR-125b regulation of the p53 network, but not individual miRNA-target pairs, is conserved

Our results reveal that *miR-125b* regulation of the p53 network is conserved at the network-level over the course of vertebrate evolution, but individual miRNA-target pairs are evolving rapidly.

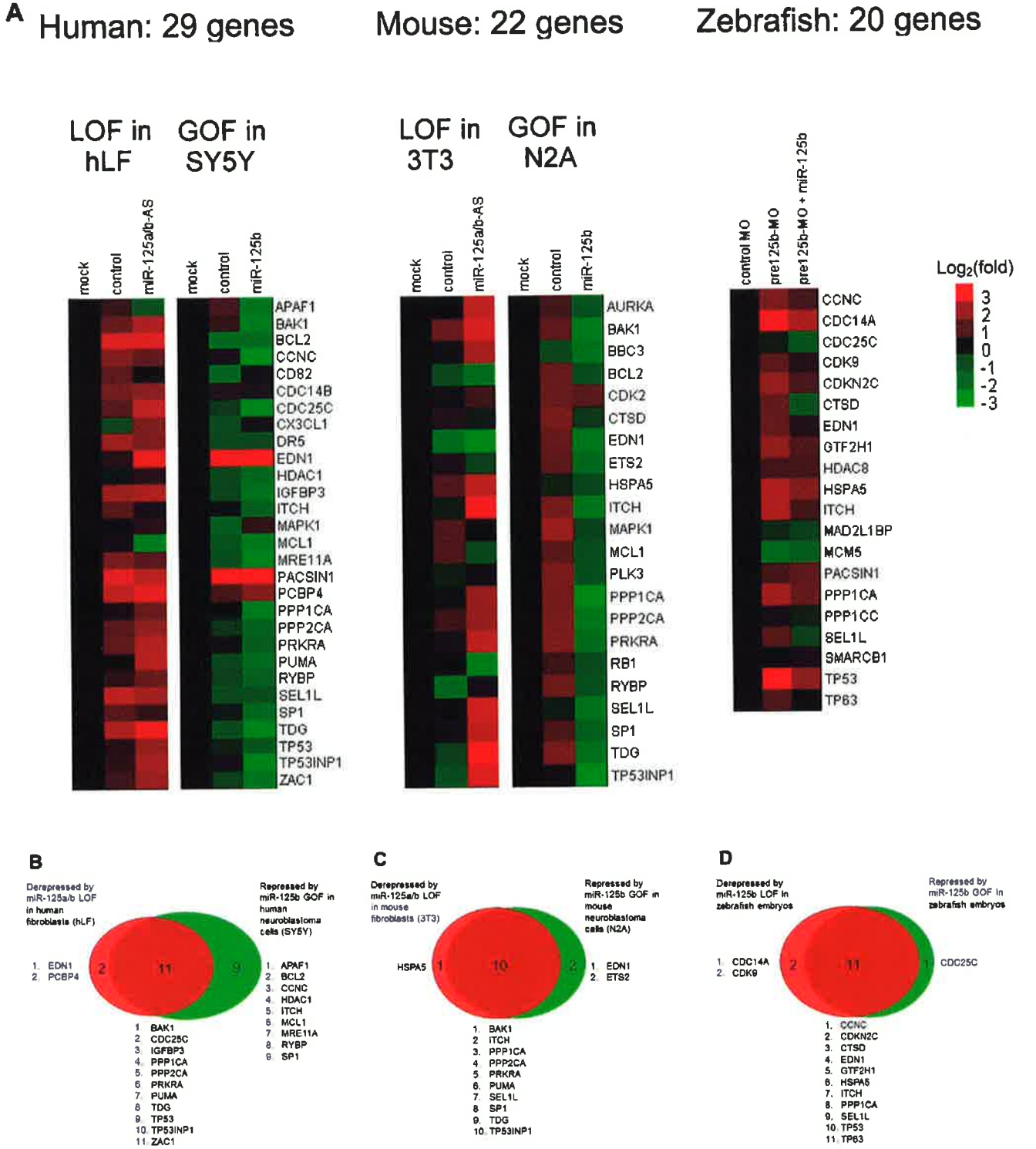


Figure 2. GOF/LOF screen for p53 network genes regulated by miR-125b. (A) Loss-of-function (LOF) screens were performed in human primary lung fibroblasts (hLF) or mouse 3T3 fibroblasts by transfecting an antisense RNA against both miR-125a and miR-125b (miR-125a/b-AS), or by microinjecting morpholinos (MO) against pre-*mir-125b* hairpin precursors (all 3 isoforms) into zebrafish embryos. Gain-of-function (GOF) screens were performed in human SH-SY5Y and mouse N2A neuroblastoma by transfecting the miR-125b duplex into cells in culture, or by coinjecting the miR-125b duplex with the morpholinos against pre-*mir-125b* into zebrafish embryos. Fold changes in gene expression were measured by qRT-PCR twenty-four hours after transfection or injection, relative to the mock and negative control miRNA or morpholino, and shown as $\log_2$(fold change) using a heat-map. (B) Human: 13 genes were significantly derepressed by a loss of miR-125b, while 20 genes were significantly repressed by a gain of miR-125b, making a total of 22 genes that passed the screen ($P<0.05$, fold change > 1.3, relative to mock control). (C) Mouse: 11 genes were significantly derepressed by a loss of miR-125b, while 12 genes were significantly repressed by a loss of miR-125b, making a total of 13 genes that passed the screen ($P<0.05$, fold change > 1.3, relative to mock control). (D) Zebrafish: 13 genes were significantly derepressed by a loss of pre-*mir-125b* ($P<0.05$, fold change > 1.3, relative to control MO), while 12 genes were significantly repressed/rescued by a gain of miR-125b ($P<0.05$, fold change > 1.3, relative to pre-*mir-125b* MO), making a total of 14 genes that passed the screen. All experiments were performed with at least three biological replicates.
doi:10.1371/journal.pgen.1002242.g002

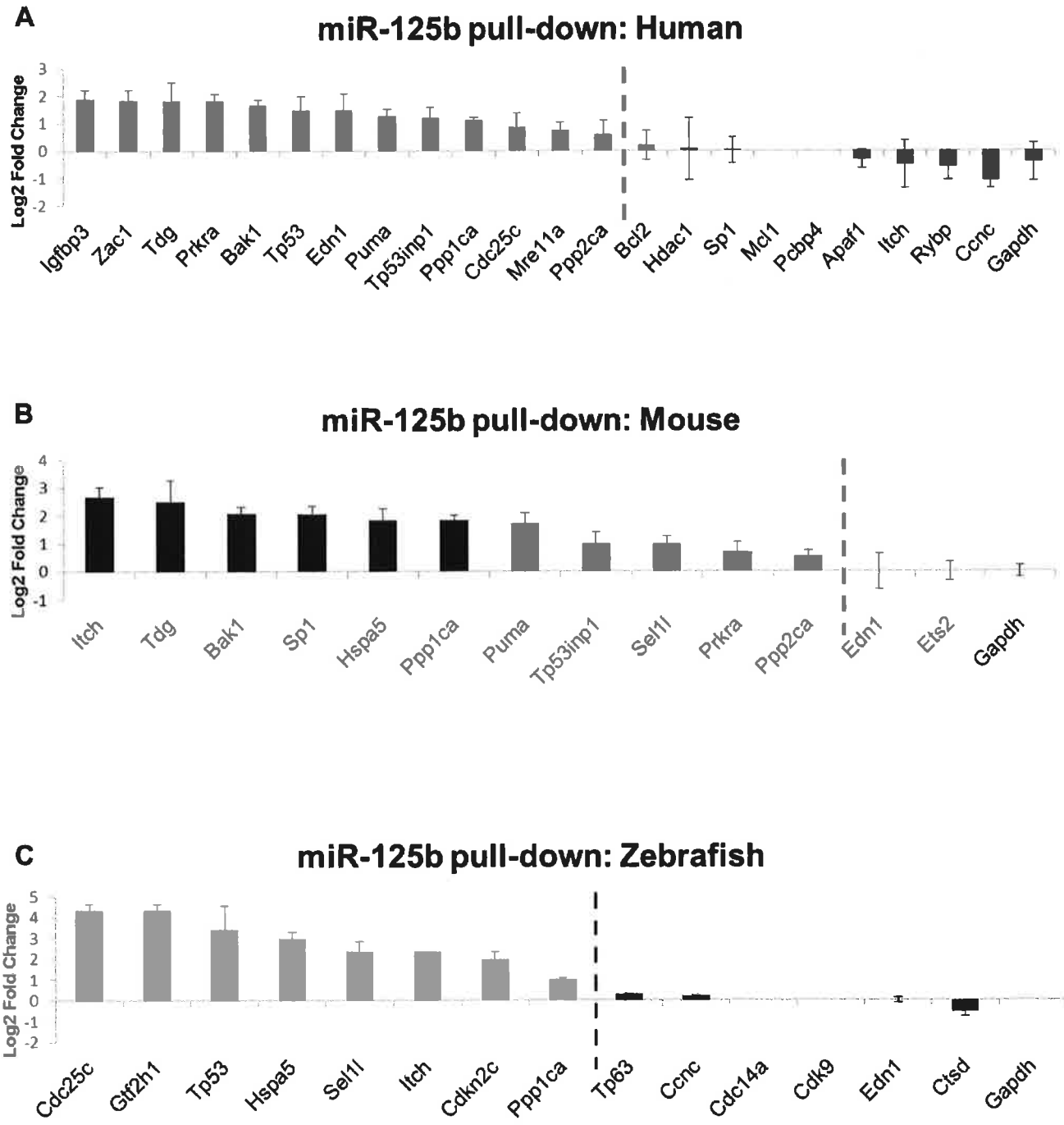


Figure 3. Direct binding of miR-125b to p53 network targets. Biotinylated miR-125b was used as bait to pull-down mRNAs bound to miR-125b, using streptavidin-conjugated magnetic beads. The mRNAs were quantified by qRT-PCR, normalized to *Gapdh,* and then compared relative to the same mRNA species pulled down by a biotinylated *C. elegans* negative control microRNA. The enrichment of mRNAs bound to miR-125b is presented as mean $\log_2$ fold change ± s.e.m. ($n \geq 3$ biological replicates). (A) Human: 13 out of 22 candidate targets were significantly enriched by miR-125b pull-down in human primary lung fibroblasts (hLF) 24 hours after transfection. (B) Mouse: 11 out of 13 candidate targets were significantly enriched by miR-125b pull-down in mouse 3T3 fibroblasts 24 hours after transfection. (C) Zebrafish: 8 of 14 candidate targets were significantly enriched by miR-125b pull-down in zebrafish embryos 24 hours after injection. Dashed line: cutoff for genes that were significantly enriched (Log_2 Fold change > 0.5, $P<0.05$).
doi:10.1371/journal.pgen.1002242.g003

To summarize our results, our list of predicted *miR-125b* targets in the p53 network (Table S1) was filtered and reclassified according to the results of the screen and validation assays (Figure 5). From the GOF/LOF screen we were able to identify mRNAs perturbed by miR-125b. However these results did not discriminate between direct or indirect targets. To supplement these experiments the pull-down assay was used to uncover mRNAs physically associated with *miR-125b*. Of note, the pull-down might not identify mRNA targets that are rapidly degraded, and as such the luciferase reporter assay can complement its shortcomings. Taken together

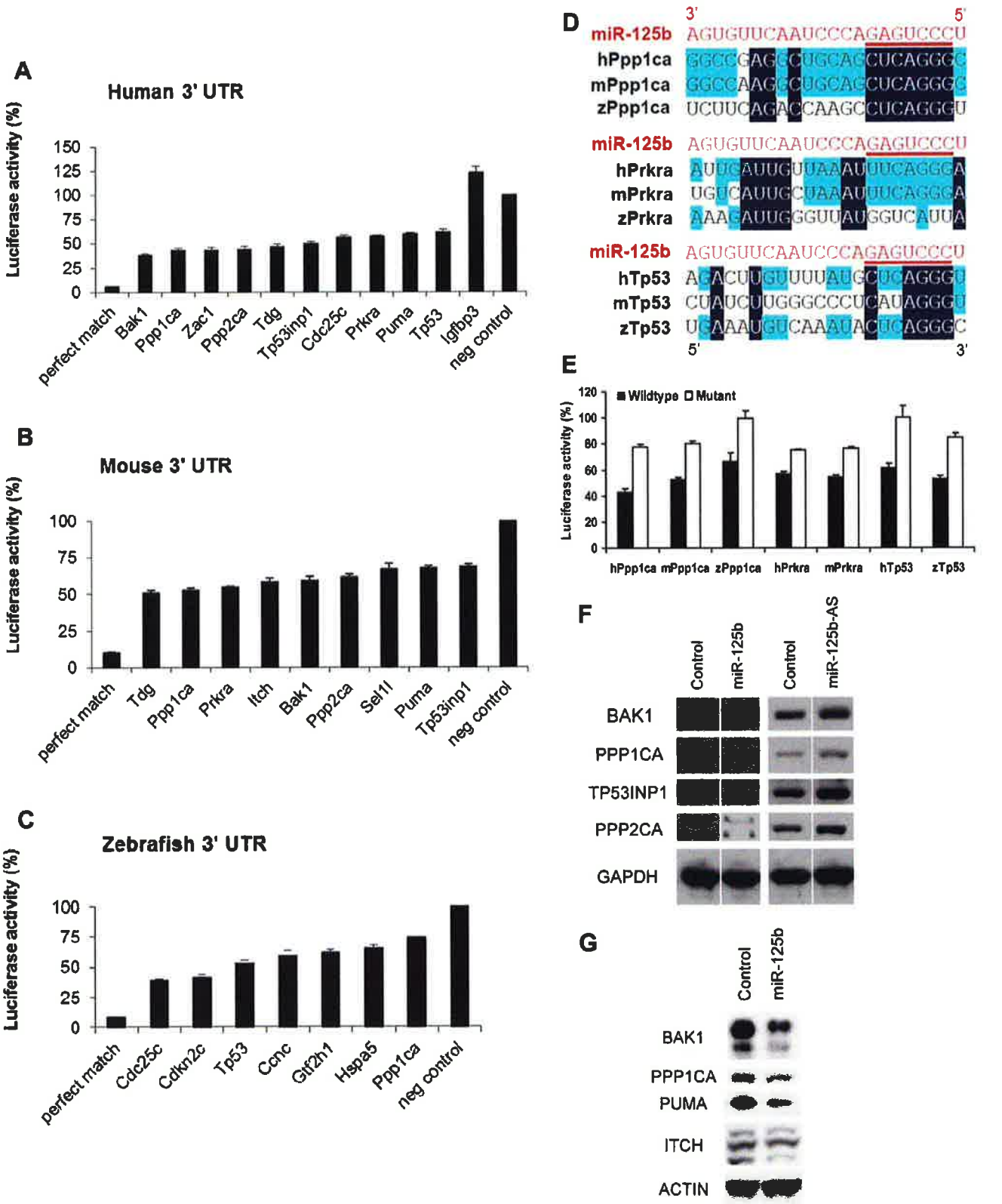


Figure 4. Validation of miR-125b targets. Candidate p53 network genes that were positive in both the GOF/LOF screen and miR-125b pull-down were validated for targeting by miR-125b using the 3′ UTR luciferase reporter assay and Western blots for protein expression. (A-C), Reporter genes containing the full-length 3′ UTRs of each selected target gene were co-transfected with miR-125b duplex into 293T cells. Luciferase readings were obtained 48 hours after transfection and presented here as the average percentage of luciferase activity ± s.e.m. ($n \geq 3$) relative to a scrambled duplex co-transfected control (100%). A reporter containing a 23-nucleotide-binding-site with perfect complementarity to miR-125b was used as the perfect match positive control, while the unmodified luciferase reporter was used as the empty negative control. (A) Human: 10 out of 13 candidate genes' 3′ UTRs showed significant repression by miR-125b relative to the control ($p<0.01$). (B) Mouse: 9 out of 11 candidate genes' 3′ UTRs showed

significant repression by miR-125b relative to the control ($p<0.01$). (C) Zebrafish: 7 out of 8 candidate genes' 3′ UTRs showed significant repression by miR-125b relative to the control ($p<0.01$). (D) Alignment of predicted *miR-125b* binding sites in the 3′UTRs of Ppp1ca, Prkra and Tp53 across three species. Seed-binding sequences are underlined. Bases conserved in two (blue) or three (black) species are highlighted. (E) The 3′UTR seed-binding sequences of 7 target mRNAs were mutated and assayed for direct binding to *miR-125b* using the luciferase reporter assay, relative to wild-type 3′UTR sequences. (E) The seed-binding sequences in the 3′UTR of 7 predicted target mRNAs were mutated and compared to wild-type sequences for binding to *miR-125b* using luciferase reporter assay. (F-G) Western blot analysis of protein expression of selected target genes two days after a transfection of miR-125b duplex, miR-125b antisense (AS) or negative control duplex or negative control antisense. (F) Western blots showed that miR-125b repressed BAK1, PPP1CA, TP53, and PPP2CA levels in human SH-SY5Y neuroblastoma cells, while the antisense RNA miR-125b-AS derepressed expression of these proteins in human ReNcell VM neural progenitor cells. (G) Western blots showed that miR-125b repressed BAK1, PPP1CA, PUMA, and ITCH levels in mouse N2A neuroblastoma cells. Abbreviations: h, human; m, mouse; z, zebrafish.
doi:10.1371/journal.pgen.1002242.g004

the three assays provide a powerful means to identify direct *miR-125b* targets. In order to minimize false positives, we counted the number of assays for which each gene target was positive, and gene targets that failed to pass at least 2 assays in at least one vertebrate species were filtered out. Predicted targets that passed 3 assays (red), 2 assays (orange), 1 assay (yellow), or predicted targets that failed all assays but whose orthologues in other species passed 3 assays of direct regulation by *miR-125b* (pink), were colored as indicated (Figure 5). Using our conservative estimate of *miR-125b* targets in the p53 network, we found that in all three vertebrates we examined – humans, mice and zebrafish – *miR-125b* regulates multiple p53 network genes. This shows that *miR-125b* regulation of the p53 network is conserved at least at the network level. However very few individual gene targets of *miR-125b* in the p53 network were conserved across all three vertebrates (Figure 5; Figure 6A-6C). Instead, conserved *miR-125b* regulation of the p53 network appears to occur through evolving miRNA-target pairs in the three vertebrates – zebrafish (Figure 6A), mouse (Figure 6B), and humans (Figure 6C). In general, we observe *miR-125b* regulating 2 general classes of genes in the p53 network: (i) apoptosis regulators like *Bak1*, *Igfbp3*, *Itch*, *Puma*, *Prkra*, *Tp53inp1*, *Tp53*, and *Zac1*, and (ii) cell-cycle regulators like *cyclin C*, *Cdc25c*, *Cdkn2c*, *Edn1*, *Ppp1ca*, and *Sel1l*.

Because *miR-125b* represses both pro-apoptosis and anti-apoptosis genes, as well as both proliferation and cell-cycle arrest genes in all three vertebrates (Figure 5), *miR-125b* appears to modulate the p53 network on the whole through an incoherent feedforward loop (FFL) [33,34] acting on the cellular processes of apoptosis and cell proliferation (Figure 6D). An incoherent type-2 FFL is a regulatory pattern in which X represses a target Z and also represses Y, another repressor of Z (Figure 6D). Incoherent FFLs have been found in the transcription factor networks of human embryonic stem cells and hematopoietic stem cells, and have been shown to modulate E2F1 dosage in the Myc-E2F1 pathway [35-37]. Besides accelerating responses and acting as amplitude filters [38-40], the incoherent FFL motif is also a noise buffering motif that reduces the variance of network dosage [41–43]. Thus our finding that incoherent FFLs fit the overall structure of network relationships between *miR-125b* and the p53-mediated processes, suggests that *miR-125b* is fine-tuning and buffering p53 network dosage.

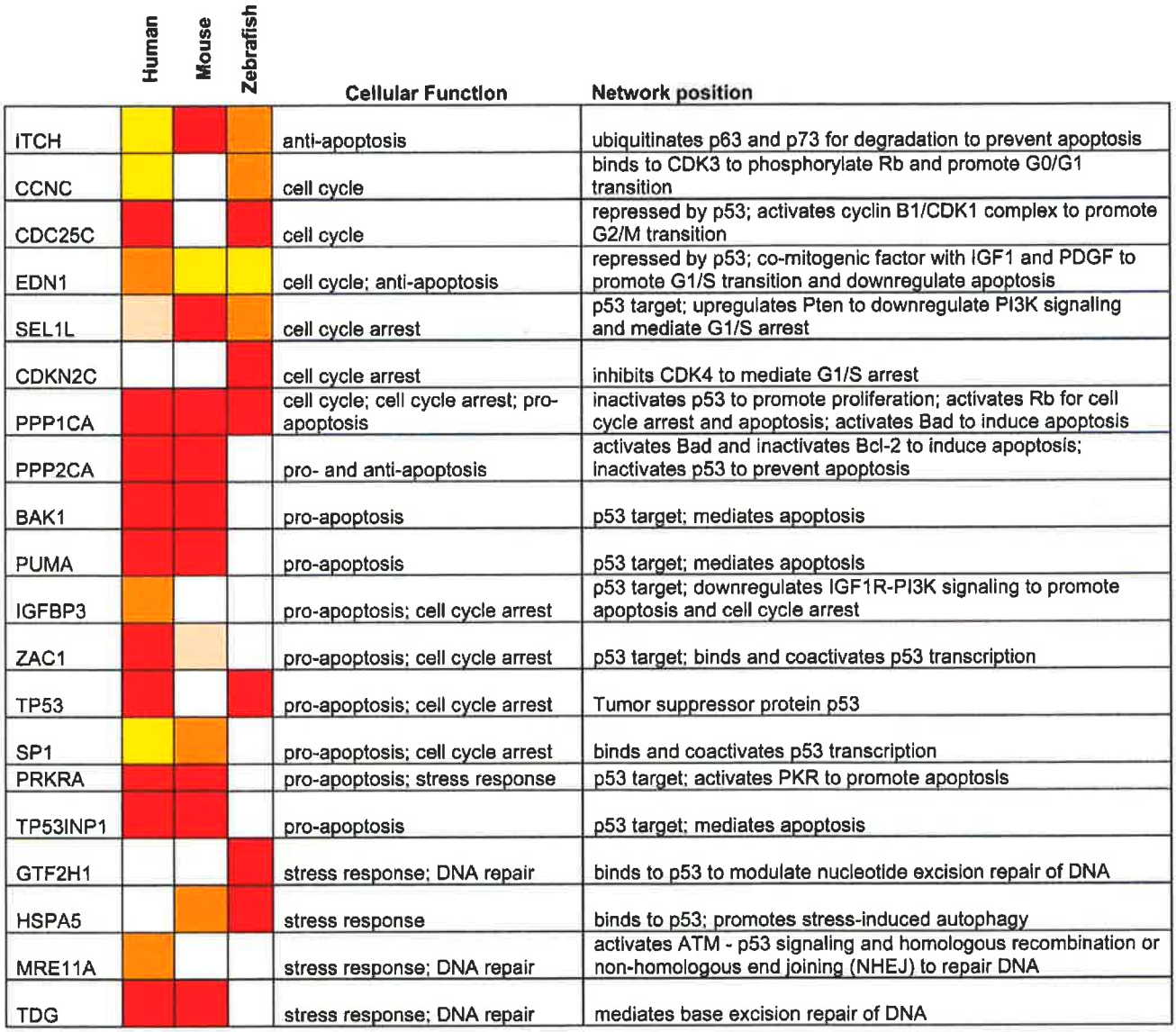
Discussion

In this study, we sought to identify direct targets of *miR-125b* in the p53 network of humans, mice and zebrafish, to better understand how *miR-125b* regulates the p53 network throughout evolution and how that might relate to its conserved role in regulating tissue stem cells.

We identified 20 direct targets of *miR-125b* in the p53 network, including 15 novel targets like *Zac1*, *Puma*, *Itch* and *Cdc25c*, and also targets like *Bak1* and *Tp53* that were identified in previous studies [9,16–18]. In general, we found that *miR-125b* directly represses 2 classes of genes: apoptosis regulators and cell-cycle regulators. With the exception of *Ppp1ca*, *Itch* and *Edn1*, very few individual targets were strictly conserved throughout vertebrate evolution. Instead, we found that only the network-level of regulation was conserved, and *miR-125b*-regulation of individual apoptosis and proliferation regulators appears to be evolving rapidly from species to species. This observation suggests that, at least within the vertebrates, the 3′ UTR sequences of each gene target is evolving rapidly via neutral genetic drift. In other words, the loss or gain of a single *miR-125b*-binding site in the 3′ UTR of most genes appears to have a relatively insignificant effect on the fitness of an organism. On the other hand, the strict conservation of *miR-125b*-regulation at the network-level in humans, mice and zebrafish, suggests that natural selection acts on the network-level rather than the gene-level with regard to miRNA-target evolution. It will be interesting to see if this novel paradigm applies to other microRNAs or gene networks as well.

Previous studies on miRNA evolution have suggested that a relatively poor conservation of individual miRNA-target pairs but strong conservation of a miRNA-gene network relationship is consistent with miRNAs' role as buffers of gene expression [42,44,45]. Our observation that an incoherent FFL-like network motif fits the overall structure of the *miR-125b* - p53 network models with respect to apoptosis and cell proliferation, lends further support to this idea since incoherent FFL network motifs are well-adapted for noise filtering [41,43,46]. It is thought that miRNAs are at least partially responsible for the phenomenon of developmental or phenotypic stability within each species [41,42], termed "canalization" by C. H. Waddington [47]. These studies suggest that miRNAs have a conserved role in regulating the overall stability of pathways/networks, a role which is relatively unaffected by the loss or gain of individual miRNA-targets over the course of evolution. A network buffering function has also been suggested for the regulation of muscle development by *miR-1* throughout evolution, regulation of the Wnt pathway by *miR-8*, and fine-regulation of Pten dosage by a variety of miRNAs [48–50]. Our findings suggest that the fine-tuning of p53 network dosage by *miR-125b* is another example of this paradigm.

Fine-regulation of p53 network dosage by *miR-125b* may also explain *miR-125b*'s conserved role in regulating tissue stem cell homeostasis. In *C. elegans*, loss-of-function mutations in *lin-4* lead to a delay in differentiation and thus expansion of vulval precursor cells, seam stem cells in the lateral hypodermis and mesoblasts, causing multiple defects in larval development [6]. In zebrafish, loss of *miR-125b* leads to widespread p53-dependent apoptosis with consequent defects in early embryogenesis, especially in neurogenesis and somitogenesis [16]. Overexpression of *miR-125a/b* causes an expansion of mammalian hematopoietic stem cells (HSCs) and aberrant differentiation, leading to myeloid leukemia [9,10] and also lymphoid leukemia if *miR-125b* is overexpressed in fetal liver HSC-enriched cells [12]. However, the molecular underpinnings of *miR-125a/b*'s regula-



	Human	Mouse	Zebrafish	Cellular Function	Network position
ITCH				anti-apoptosis	ubiquitinates p63 and p73 for degradation to prevent apoptosis
CCNC				cell cycle	binds to CDK3 to phosphorylate Rb and promote G0/G1 transition
CDC25C				cell cycle	repressed by p53; activates cyclin B1/CDK1 complex to promote G2/M transition
EDN1				cell cycle; anti-apoptosis	repressed by p53; co-mitogenic factor with IGF1 and PDGF to promote G1/S transition and downregulate apoptosis
SEL1L				cell cycle arrest	p53 target; upregulates Pten to downregulate PI3K signaling and mediate G1/S arrest
CDKN2C				cell cycle arrest	inhibits CDK4 to mediate G1/S arrest
PPP1CA				cell cycle; cell cycle arrest; pro-apoptosis	inactivates p53 to promote proliferation; activates Rb for cell cycle arrest and apoptosis; activates Bad to induce apoptosis
PPP2CA				pro- and anti-apoptosis	activates Bad and inactivates Bcl-2 to induce apoptosis; inactivates p53 to prevent apoptosis
BAK1				pro-apoptosis	p53 target; mediates apoptosis
PUMA				pro-apoptosis	p53 target; mediates apoptosis
IGFBP3				pro-apoptosis; cell cycle arrest	p53 target; downregulates IGF1R-PI3K signaling to promote apoptosis and cell cycle arrest
ZAC1				pro-apoptosis; cell cycle arrest	p53 target; binds and coactivates p53 transcription
TP53				pro-apoptosis; cell cycle arrest	Tumor suppressor protein p53
SP1				pro-apoptosis; cell cycle arrest	binds and coactivates p53 transcription
PRKRA				pro-apoptosis; stress response	p53 target; activates PKR to promote apoptosis
TP53INP1				pro-apoptosis	p53 target; mediates apoptosis
GTF2H1				stress response; DNA repair	binds to p53 to modulate nucleotide excision repair of DNA
HSPA5				stress response	binds to p53; promotes stress-induced autophagy
MRE11A				stress response; DNA repair	activates ATM - p53 signaling and homologous recombination or non-homologous end joining (NHEJ) to repair DNA
TDG				stress response; DNA repair	mediates base excision repair of DNA

Legend

No. of Screen/Validation assays that each target passed	
3	GOF/LOF, pull-down & luciferase
2	GOF/LOF, & pull-down or luciferase
1	GOF/LOF
0	predicted target; validated in other species

Figure 5. Summary of genes in p53 network that are directly targeted by miR-125b. Only targets that passed ≥ 2 validation assays, in at least one species, are shown. Red: predicted targets validated by 3 assays; Orange: predicted targets validated by 2 assays; Yellow: predicted targets validated by 1 assay; Pink: predicted targets not validated by any assay, but validated by 3 assays in another species.
doi:10.1371/journal.pgen.1002242.g005

tion of tissue stem cell homeostasis had remained unclear largely due to the complex nature of microRNA regulation of gene networks. The 2 classes of *miR-125b* targets in the p53 network, and the incoherent FFL network motifs that we found, may at least partially explain how *miR-125b* regulates tissue stem cells in vertebrates. By fine-tuning both apoptosis regulators and cell-cycle regulators, *miR-125b* may fine-tune the p53 network dosage to drive the self-renewal of tissue stem cells. It could explain how overexpression of *miR-125b* leads to an expansion of self-renewing hematopoietic stem cells while loss of *miR-125b* leads to aberrant apoptosis and proliferation, with consequent defects in tissue differentiation.

Several studies have implicated *miR-125b* as an oncogene in a variety of mammalian tissue compartments, e.g. leukemia, neuroblastoma, prostate cancer and breast cancer [9–18]. These studies have ascribed *miR-125b*'s anti-apoptotic effect as an oncogene to its direct suppression of Bak1 or Tp53 [9,16-18]. On the other hand, several research groups have also reported *miR-125b*'s role as a

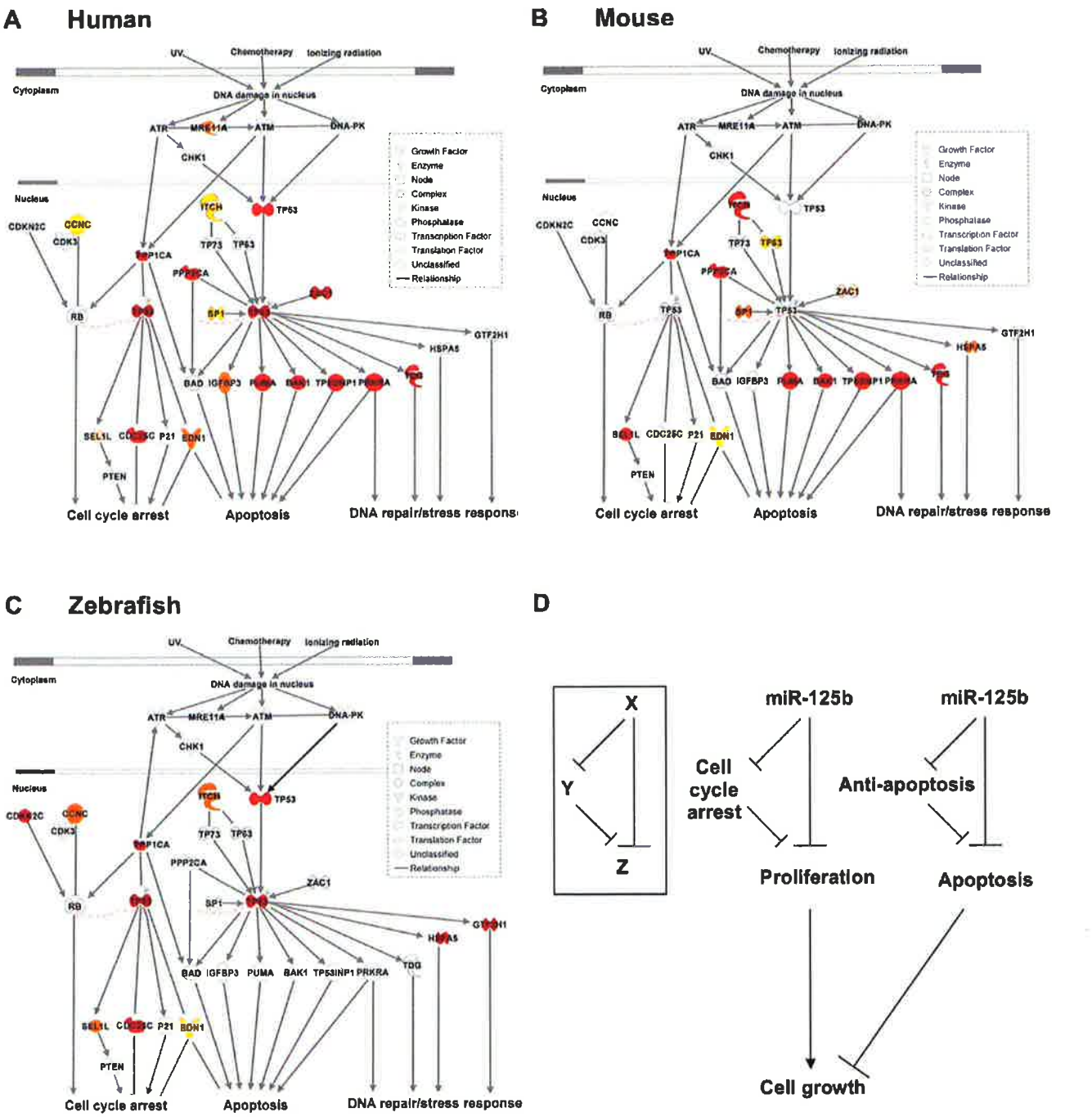


Figure 6. Models of miR-125b regulation of p53 networks in humans, mice, and zebrafish. (A) Human p53 network. (B) Mouse p53 network. (C) Zebrafish p53 network. Models were constructed by Ingenuity Pathway Analysis. Red: predicted targets validated by 3 assays; Orange: predicted targets validated by 2 assays; Yellow: predicted targets validated by 1 assay; Pink: predicted targets not validated by any assay, but validated by 3 assays in another species. (D) Incoherent feedforward loop (FFL) motifs characterize miR-125b regulation of p53 network genes that mediate apoptosis or cell cycle arrest.
doi:10.1371/journal.pgen.1002242.g006

potential tumor suppressor by suppressing proliferation in cell-culture models [19–23]. Our identification of 20 direct targets of *miR-125b* in the p53 network reconciles these findings because *miR-125b* modulates the expression of both apoptosis regulators and cell-cycle regulators. Although *miR-125b*'s suppression of p53 itself is not conserved in mice, *miR-125b*'s anti-apoptotic role – through suppression of multiple pro-apoptosis regulators in the p53 network – appears to be conserved in vertebrates. *miR-125b*'s ability to fine-tune the subtle balance of apoptosis vs. cell-cycle regulators and thus buffer the p53 network dosage in different contexts, could explain why *miR-125b* dysregulation can lead to either tumor suppression or oncogenesis depending on the context. It is possible that this buffering feature of *miR-125b* represents a general principle of miRNA regulation of gene networks.

Materials and Methods

Prediction of miR-125b targets in the p53 network

A list of p53-associated genes was compiled from the p53 Knowledgebase website [30] and from the Ingenuity Pathway Analysis™ database [29]. The targets of miR-125b in human and mouse were predicted by TargetScan [31]. The targets of miR-125b in zebrafish were predicted by MicroCosm [32]. The human homologues of mouse and zebrafish targets were identified by the DAVID gene ID conversion tool.

Cell culture and transfection

Human lung fibroblast cells, human neuroblastoma SH-SY5Y cells, mouse neuroblastoma Neuro-2A cells, mouse fibroblast Swiss-3T3 cells and human HEK-293T cells were maintained in DMEM media, supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin (Invitrogen). Neuro-2A cells, 3T3 cells, SH-SY5Y cells and human lung fibroblast cells were transfected in suspension with 5×10^5 cells per well in 6-well plates using lipofectamin-2000 (Invitrogen). miRNA duplexes and antisense oligonucleotides (Ambion) were transfected at a final concentration of 80 nM.

Microinjection in zebrafish embryos

Wild-type zebrafish were maintained by standard protocols [51]. All injections were carried out at 1–4 cell stage with 2 nl of solution into each embryo. In the knockdown experiments, miR-125b morpholinos were injected at 0.75 pmole/embryo (lp125bMO1/2/3 indicates the co-injection of three lp125bMOs, 0.25 pmole each); miR-125b duplex was injected as 37.5 fmole/embryo.

Quantitative RT-PCR

RNA was extracted from cells or zebrafish embryos using Trizol reagent (Invitrogen) and subsequently column-purified with RNeasy kits (Qiagen). For qRT-PCR of miR-125a, miR-125b and RNU6B, 100 ng of total RNA was reverse-transcribed and subjected to Taqman microRNA assay (Applied Biosystems). For qRT-PCR of mRNAs, cDNA synthesis was performed with 1 μg of total RNA using the High Capacity cDNA Archive Kit (Applied Biosystems). The expression of all genes was analyzed by SYBR assay using the Applied Biosystems real-time PCR system or the Fluidigm 48x48 dynamic array system (Fluidigm) following the manufacturer's protocol.

miRNA–target pull-down assay

50 ul of streptavidin coated magnetic beads (Invitrogen) were blocked with 1 mg/ml yeast tRNA and 1 mg/ml BSA in 1 ml lysis buffer (20 mM Tris pH 7.5, 100 mM KCl, 5 mM MgCl2, 0.3% NP-40) for 2 hours at 4°C and wash twice with lysis buffer. hsa-miR-125b or cel-miR-67 (negative control) duplex was synthesized with a biotin conjugated at the 3′ end of the active strand by Dharmacon Research Inc. The miRNAs were transfected into human lung fibroblasts or mouse 3T3 fibroblasts at a final concentration of 80 nM as described above. The miRNAs were also injected into zebrafish embryos at 1 to 4-cell stage at a final concentration of 37.5 fmole/embryo. After 24 hours, cells from 3 wells of fibroblasts or 50 zebrafish embryos were incubated with 500 ul cold lysis buffer containing freshly added 100 units/ml RNase inhibitor (Invitrogen) and protease inhibitor cocktail (Roche) for 20 minutes on ice. After the cell debris is removed by centrifugation, the lysate was incubated with pre-blocked streptavidin coated beads for 2 hours at 4°C. Subsequently, the beads were washed 5 times with cold lysis buffer and incubated with Trizol for RNA extraction.

Luciferase reporter assay

The whole 3′ UTR of target genes were cloned into the psiCHECK-2 vector (Promega), between the XhoI and NotI site, immediately 3' downstream of the *Renilla* luciferase gene. For selected targets, we introduced 3 point mutations into the 7-nt seed-binding sequence using inverse PCR with non-overlapping primers carrying the mutated sequences. 10 ng of each psi-CHECK-2 construct was co-transfected with 10 nM miRNA duplexes or into HEK-293T cells in a 96-well plate using lipofectamin-2000 (Invitrogen). After 48 hours, the cell extract was obtained; Firefly and *Renilla* luciferase activities were measured with the Dual-Luciferase reporter system (Promega) according to the manufacturer's instructions.

Western blot assay

Cells were lysed in RIPA buffer (Pierce). Proteins were separated by a 10% polyacrylamide gel and transferred to a methanol-activated PVDF membrane (GE Healthcare). The membrane was blocked for one hour in PBST containing 7.5% milk and subsequently probed with primary antibodies (Santa Cruz) overnight at 4°C. After 1-hour incubation with goat-anti-mouse HRP-conjugated secondary antibody (Santa Cruz), the protein level was detected with luminol reagent (Santa Cruz).

Statistical analysis

Two-tail T-tests were used to determine the significance of differences between the treated samples and the controls where values were obtained from luciferase reporter assay or qRT-PCR. The tests were performed using Microsoft Excel where the test type is always set to two-sample equal variance.

Supporting Information

Figure S1 Mature miR-125b levels before and after overexpression or knockdown. (A) The level of miR-125b in human SH-SY5Y cells one day after a transfection with mock (lipofectamin2000 only), negative control duplex (NC-DP) or miR-125b duplex (125b-DP). (B) The level of miR-125b in mouse N2A cells one day after a transfection with mock, NC-DP or 125b-DP. (C) The level of miR-125a and miR-125b in human lung fibroblasts one day after a transfection with mock, negative control antisense (NC-AS) or miR-125a antisense and miR-125b antisense cotransfection (125ab-AS). (D) The level of miR-125a and miR-125b in mouse SWISS-3T3 fibroblasts one day after a transfection with mock, NC-AS or (125ab-AS). In all panels, the levels of miR-125a and miR-125b were quantified by real-time PCR, and presented as $\log_2$ (fold change) ± s.e.m. ($n\geq3$) relative to the levelsof RNU6B loading control.
(TIF)

Table S1 Genes in p53 network with predicted miR-125b binding sites. [a] Hsa: *Homo sapiens*, humans. [b] Mmu: *Mus musculus*, mice. [c] Dre: *Danio rerio*, zebrafish. [d] Non-official but common gene name that is used in this paper.
(PDF)

Acknowledgments

The authors thank all our colleagues in Singapore and Boston, especially Huangming Xie, Beiyan Zhou, and Moonyoung Um for their contribution to the initiation of the project; Kian Leong Lee, Guo Guoji, Hao Zhu, Wai Leong Tam, Senthil Raja Jayapal, Marshall Thomas, and Ricardo

Henriques for help and fruitful discussions; Soh Boon Seng for providing the human lung fibroblasts and SH-SY5Y cells; Wa Xian and Frank McKeon for providing Swiss-3T3 cells; Lim Qing En and Too Heng Phon for providing Neuro-2A cells; Michael Kharas for help with flow cytometry.

References

1. Lee RC, Feinbaum RL, Ambros V (1993) The C-elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75: 843–854.
2. Bartel DP, Chen CZ (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nature Reviews Genetics 5: 396–400.
3. He L, Hannon GJ (2004) MicroRNAs: Small RNAs with a big role in gene regulation (vol 5, pg 522 2004). Nature Reviews Genetics 5: 522-+.
4. Inui M, Martello G, Piccolo S (2010) MicroRNA control of signal transduction. Nature Reviews Molecular Cell Biology 11: 252–263.
5. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, et al. (2002) Identification of tissue-specific microRNAs from mouse. Current Biology 12: 735–739.
6. Chalfie M, Horvitz HR, Sulston JE (1981) Mutations that lead to reiterations in the cell lineages of C-elegans. Cell 24: 59–69.
7. Le MTN, Xie HM, Zhou BY, Chia PH, Rizk P, et al. (2009) MicroRNA-125b Promotes Neuronal Differentiation in Human Cells by Repressing Multiple Targets. Molecular and Cellular Biology 29: 5290–5305.
8. Rybak A, Fuchs H, Smirnova L, Brandt C, Pohl EE, et al. (2008) A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nature Cell Biology 10: 987–993.
9. Guo SQ, Lu J, Schlanger R, Zhang H, Wang JY, et al. (2010) MicroRNA miR-125a controls hematopoietic stem cell number. Proceedings of the National Academy of Sciences of the United States of America 107: 14229–14234.
10. O'Connell RM, Chaudhuri AA, Rao DS, Gibson WSJ, Balazs AB, et al. (2010) MicroRNAs enriched in hematopoietic stem cells differentially regulate long-term hematopoietic output. Proceedings of the National Academy of Sciences of the United States of America 107: 14235–14240.
11. Bousquet M, Quelen C, Rosati R, Mansat-De Mas R, La Starza R, et al. (2008) Myeloid cell differentiation arrest by miR-125b-1 in myelodysplasic syndrome and acute myeloid leukemia with the t(2;11)(p21;q23) translocation. Journal of Experimental Medicine 205: 2499–2506.
12. Bousquet M, Harris MH, Zhou B, Lodish HF (2010) MicroRNA miR-125b causes leukemia. Proceedings of the National Academy of Sciences of the United States of America 107: 21558–21563.
13. Chapiro E, Russell LJ, Struski S, Cave H, Radford-Weiss I, et al. (2010) A new recurrent translocation t(11;14)(q24;q32) involving IGH@ and miR-125b-1 in B-cell progenitor acute lymphoblastic leukemia. Leukemia 24: 1362–1364.
14. Klusmann JH, Li Z, Bohmer K, Maroz A, Koch ML, et al. (2010) miR-125b-2 is a potential oncomiR on human chromosome 21 in megakaryoblastic leukemia. Genes & Development 24: 478–490.
15. Sonoki T, Iwanaga E, Mitsuya H, Asou N (2005) Insertion of microRNA-125b-1, a human homologue of lin-4, into a rearranged immunoglobulin heavy chain gene locus in a patient with precursor B-cell acute lymphoblastic leukemia. Leukemia 19: 2009–2010.
16. Le MTN, Teh C, Shyh-Chang N, Xie HM, Zhou BY, et al. (2009) MicroRNA-125b is a novel negative regulator of p53. Genes & Development 23: 862–876.
17. Zhou M, Liu ZX, Zhao YH, Ding Y, Liu H, et al. (2010) MicroRNA-125b Confers the Resistance of Breast Cancer Cells to Paclitaxel through Suppression of Pro-apoptotic Bcl-2 Antagonist Killer 1 (Bak1) Expression. Journal of Biological Chemistry 285: 21496–21507.
18. Shi XB, Xue L, Yang J, Ma AH, Zhao J, et al. (2007) An androgen-regulated miRNA suppresses Bak1 expression and induces androgen-independent growth of prostate cancer cells. Proceedings of the National Academy of Sciences of the United States of America 104: 19983–19988.
19. Scott GK, Goga A, Bhaumik D, Berger CE, Sullivan CS, et al. (2007) Coordinate suppression of ERBB2 and ERBB3 by enforced expression of micro-RNA miR-125a or miR-125b. Journal of Biological Chemistry 282: 1479–1486.
20. Shi L, Zhang JX, Pan TH, Zhou JF, Gong WY, et al. (2010) MiR-125b is critical for the suppression of human U251 glioma stem cell proliferation. Brain Research 1312: 120–126.
21. Xia HF, He TZ, Liu CM, Cui Y, Song PP, et al. (2009) MiR-125b Expression Affects the Proliferation and Apoptosis of Human Glioma Cells by Targeting Bmf. Cellular Physiology and Biochemistry 23: 347–358.
22. Hofmann MH, Heinrich J, Radziwil G, Moelling K (2009) A Short Hairpin DNA Analogous to miR-125b Inhibits C-Raf Expression, Proliferation, and Survival of Breast Cancer Cells. Molecular Cancer Research 7: 1635–1644.
23. Mizuno Y, Yagi K, Tokuzawa Y, Kanesaki-Yatsuka Y, Suda T, et al. (2008) miR-125b inhibits osteoblastic differentiation by down-regulation of cell proliferation. Biochemical and Biophysical Research Communications 368: 267–272.
24. Viswanathan S, Powers J, Einhorn W, Hoshida Y, Ng T, et al. (2009) Lin28 promotes transformation and is associated with advanced human malignancies. Nature Genetics 41: 843–848.

Author Contributions

Conceived and designed the experiments: MTNL NS-C HFL BL. Performed the experiments: MTNL NS-C SLK LC JT CT. Analyzed the data: MTNL NS-C. Contributed reagents/materials/analysis tools: VK EO AL HY. Wrote the paper: NS-C MTNL HFL BL JL.

25. Levine AJ (1997) p53, the cellular gatekeeper for growth and division. Cell 88: 323–331.
26. Agarwal ML, Taylor WR, Chernov MV, Chernova OB, Stark GR (1998) The p53 network. Journal of Biological Chemistry 273: 1–4.
27. Mayo LD, Donner DB (2002) The PTEN, Mdm2, p53 tumor suppressor-oncoprotein network. Trends in Biochemical Sciences 27: 462–467.
28. He XY, He L, Hannon GJ (2007) The guardian's little helper: MicroRNAs in the p53 tumor suppressor network. Cancer Research 67: 11099–11101.
29. Siu DC, Laurance M (2004) Rapid generations of de novo biological pathways from large-scale gene expression data using the Ingenuity Pathways Analysis application. Proceedings of the American Association for Cancer Research Annual Meeting 45: 756.
30. Lim YP, Lim TT, Chan YL, Song ACM, Yeo BH, et al. (2007) The p53 knowledgebase: an integrated information resource for p53 research. Oncogene 26: 1517–1521.
31. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15–20.
32. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Research 36: D154–D158.
33. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31: 64–68.
34. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: Simple building blocks of complex networks. Science 298: 824–827.
35. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122: 947–956.
36. Swiers G, Patient R, Loose M (2006) Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Developmental Biology 294: 525–540.
37. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT (2005) c-Myc-regulated microRNAs modulate E2F1 expression. Nature 435: 839–843.
38. Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences of the United States of America 100: 11980–11985.
39. Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. Journal of Molecular Biology 356: 1073–1081.
40. Kaplan S, Bren A, Dekel E, Alon U (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. Molecular Systems Biology 4.
41. Hornstein E, Shomron N (2006) Canalization of development by microRNAs. Nature Genetics 38: S20–S24.
42. Wu CI, Shen Y, Tang T (2009) Evolution under canalization and the dual roles of microRNAs-A hypothesis. Genome Research 19: 734–743.
43. Osella M, Bosia C, Corá D, Caselle M (2011) The Role of Incoherent MicroRNA-Mediated Feedforward Loops in Noise Buffering. PLoS Comput Biol 7: e1001101. doi:10.1371/journal.pcbi.1001101.
44. Ciliberti S, Martin OC, Wagner A (2007) Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput Biol 3: e15. doi:10.1371/journal.pcbi.0030015.
45. Wagner A (2005) Circuit topology and the evolution of robustness in two-gene circadian oscillators. Proceedings of the National Academy of Sciences of the United States of America 102: 11775–11780.
46. Tsang J, Zhu J, van Oudenaarden A (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Molecular Cell 26: 753–767.
47. Waddington CH (1959) Canalization of development and genetic assimilation of acquired characters. Nature 183: 1654–1655.
48. Stefani G, Slack FJ (2008) Small non-coding RNAs in animal development. Nature Reviews Molecular Cell Biology 9: 219–230.
49. Kennell JA, Gerin I, MacDougald OA, Cadigan KM (2008) The microRNA miR-8 is a conserved negative regulator of Wnt signaling. Proceedings of the National Academy of Sciences of the United States of America 105: 15417–15422.
50. Poliseno L, Salmena L, Zhang JW, Carver B, Haveman WJ, et al. (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465: 1033-U1090.
51. Nuesslein-Volhard C, Dahm R (2002) Zebrafish: A practical approach. Zebrafish: A practical approach: i-xviii, 1-303.

Breast Cancer RESEARCH

REVIEW

MicroRNAs and their target gene networks in breast cancer

Elizabeth O'Day and Ashish Lal*

Abstract

MicroRNAs (miRNAs) are a major class of small endogenous RNA molecules that post-transcriptionally inhibit gene expression. Many miRNAs have been implicated in several human cancers, including breast cancer. Here we describe the association between altered miRNA signatures and breast cancer tumorigenesis and metastasis. The loss of several tumor suppressor miRNAs (miR-206, miR-17-5p, miR-125a, miR-125b, miR-200, let-7, miR-34 and miR-31) and the overexpression of certain oncogenic miRNAs (miR-21, miR-155, miR-10b, miR-373 and miR-520c) have been observed in many breast cancers. The gene networks orchestrated by these miRNAs are still largely unknown, although key targets have been identified that may contribute to the disease phenotype. Here we report how the observed perturbations in miRNA expression profiles may lead to disruption of key pathways involved in breast cancer.

Introduction

MicroRNAs (miRNAs) are an evolutionarily conserved class of small, approximately 22-nucleotide non-coding RNAs that decrease gene expression post-transcriptionally in a sequence-specific manner. Most miRNAs are transcribed in the nucleus by RNA polymerase II (although some miRNAs are also transcribed by RNA polymerase III [1]) as long primary transcripts (pri-miRNAs) that undergo processing by Drosha and DGCR8, resulting in an approximately 70-nucleotide stem-loop RNA (pre-miRNA). Pre-miRNAs are subsequently exported to the cytoplasm via Exportin 5 and cleaved by Dicer, giving rise to approximately 22-nucleotide RNA duplexes. The strand with decreased base-pairing at its 5′ end is then selected to function as a mature miRNA, while the other strand (also referred to as the passenger strand) is typically degraded [2,3]. The mature miRNA associates with Argonautes and other proteins to form the RNA-induced silencing complex (miRISC), which then binds to target mRNAs via partial complementarity. Many transcription factors that regulate mRNA transcription also control miRNA biogenesis. Although transcription plays a major role in miRNA biogenesis, additional mechanisms, such as DNA methylation, can also regulate miRNA expression [4]. Mammalian miRNAs predominantly act by binding to the 3′ untranslated region (UTR) of cognate mRNAs. However, there is growing evidence that they can also downregulate the expression of some genes by base-pairing to the coding region [5-8] or the 5′ UTR of some mRNAs [9].

Initially, miRNAs were thought to function mainly by suppressing mRNA translation [10]. However, two recent studies combined proteomics and microarrays to reveal that changes in protein expression mediated by a miRNA are usually associated with altered mRNA expression, suggesting that mRNA degradation may be the major component of mammalian miRNA repression [11,12]. Although most studies suggest that miRNAs negatively regulate gene expression by base-pairing to the 3′ UTR, a few recent examples have shown that miRNAs can also upregulate the translation of their target mRNAs [13].

According to miRBase release 14 (September 2009), more than 700 human miRNAs have been identified [14]. Regardless of the mechanism by which they regulate gene expression, each miRNA can potentially regulate the expression of hundreds of genes and a single transcript can be targeted by multiple miRNAs to concurrently downregulate multiple proteins in the same pathway. In fact, almost one-third of the protein-coding genes are susceptible to miRNA regulation [15] and, as a consequence, many miRNAs seem to play pivotal roles in important biological processes, including cellular proliferation, differentiation and apoptosis [16]. Not surprisingly, aberrant miRNA expression is a hallmark of several diseases, cancer in particular [17,18]. In this review we will focus on recent advances on the functions of miRNAs in breast tumor development and metastasis.

*Correspondence: alal@idi.harvard.edu
Immune Disease Institute, Program in Cellular and Molecular Medicine, Children's Hospital Boston and Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

Finally, based on reported targets of specific breast cancer-associated miRNAs, we have built a direct gene interaction network to illustrate how the targets of these miRNAs interact with each other via protein-protein or protein-DNA interactions. This approach reveals that many genes involved in breast cancer are embedded in a miRNA network that controls their expression leading to breast cancer pathogenesis.

MicroRNAs as critical regulators of tumorigenesis

In normal cells, miRNAs control normal rates of cellular growth, proliferation, differentiation and apoptosis. Since miRNAs inhibit cell cycle progression and drive terminal differentiation, downregulation of some miRNAs may play an important role in the development or progression of cancer [19]. In this context, at least three observations early in the history of miRNAs suggested their potential role in cancer. First, the earliest miRNAs discovered in *Caenorhabditis elegans* and *Drosophila* were shown to regulate cellular proliferation and apoptosis [20,21], suggesting that deregulation of these miRNAs may lead to proliferative diseases. Second, many miRNA loci frequently map to genomic regions that are commonly amplified or deleted in various human cancers [19,22-25]. Third, tumor cell lines and malignant tumors were found to have widespread deregulation of miRNA expression [23,26,27]. Furthermore, expression of some miRNAs correlated with diverse clinicopathological parameters [23,28,29] and prognosis [30,31]. Taken together, these findings highlighted a potential role of miRNAs as new diagnostic or prognostic biomarkers. Moreover, the evidence that specific miRNAs have tumor suppressor and oncogenic functions makes them novel targets for cancer therapy.

A global decrease in miRNA expression has been observed in human cancers, suggesting that most miRNAs may act as tumor suppressors [27]. Furthermore, poorly differentiated tumors have lower miRNA levels compared to more differentiated tumors, indicating that global changes in miRNA expression may indicate the degree of cellular differentiation. In support of this hypothesis, a recent study showed that most miRNAs were expressed at lower levels in human tumor-derived cell lines compared to the corresponding tissue [26]. However, it remained unclear if the reduced miRNA expression was the cause or the consequence of malignant transformation. Remarkably, another study demonstrated that the global loss in miRNA expression in cancer promoted tumorigenesis [32]. Global loss of miRNA expression achieved by knockdown of Drosha and Dicer, two proteins essential for miRNA biogenesis, enhanced cellular growth *in vitro*. When these cells were injected into nude mice, they formed faster-growing and more invasive tumors. Taken together, these results suggest that genome-wide loss of miRNAs enhances tumorigenesis. However, it remains to be seen if the loss of all miRNAs or only a subgroup of key tumor suppressor miRNAs is the key event that promotes tumorigenesis.

Roles of microRNAs in breast cancer

Over the past few years, miRNA profiling studies have led to the identification of miRNAs that are aberrantly expressed in human breast cancer. The function of only a handful of these miRNAs in breast cancer has been investigated. As in other cancers, some miRNAs can function as tumor suppressors and other miRNAs as oncogenes. Thus, tumor formation may arise from a reduction or deletion of a tumor suppressor miRNA and/or amplification or overexpression of an oncogenic miRNA. In addition, tumor metastasis may be promoted by enhanced expression of prometastatic and/or downregulation of antimetastatic miRNAs. The functions of these miRNAs in breast tumor progression and metastasis are discussed below.

Tumor suppressor miRNAs

In this section we discuss recent findings on seven well-known miRNAs that inhibit breast tumor formation and/or metastasis (Table 1).

miR-206

A potential role of miR-206 in suppressing breast cancer was first identified by employing miRNA microarrays to compare miRNA expression between normal and breast cancer tissues [28]. miR-206 was upregulated in estrogen receptor (ER)α-negative breast cancers, suggesting a role of miR-206 in regulation of the estrogen receptor gene *ERα* (*ESR1*). Indeed, miR-206 was recently shown to inhibit the expression of *ESR1* mRNA through two binding sites in the *ESR1* 3′ UTR [33]. The latter study also showed that miR-206 expression was strongly repressed by ERα agonists, but not by an ERβ agonist or progesterone, suggesting the existence of a feedback loop. Another study showed that miR-206 expression decreased in ERα-positive human breast cancer tissues and that miR-206 suppresses ESR1 expression and inhibits growth of MCF7 breast cancer cells. In addition to miR-206, the authors found that *ESR1* mRNA was a direct target of miR-18a, miR-18b, miR-193b and miR-302c in breast cancer cells [34]. Moreover, miR-18a, miR-18b, miR-193b, miR-206 and miR-302c were also shown to induce cell cycle arrest and to inhibit estrogen-induced proliferation, with inhibition levels comparable to those achieved by *ERα* small interfering RNA (siRNA). Since dysregulation of ERα expression is a hallmark of most human breast cancers, these studies provide important insights into the molecular mechanisms that lead to

Table 1. miRNAs and their targets in breast cancer cells

miRNA	Target	Functional pathway	Reference
Tumor suppressor miRNAs			
miR-206	*ESR1*	ER signaling	[28,33]
miR-17-5p	*AIB1, CCND1, E2F1*	Proliferation	[37,38]
miR-125a, b	*HER2, HER3*	Anchorage-dependent growth	[40,41]
miR-200	*BMI1, ZEB1, ZEB2*	TGF-β signalling	[47,52,54]
let-7	*H-RAS, HMGA2, LIN28, PEBP1*	Proliferation, differentiation	[73-75]
miR-34a	*CCND1, CDK6, E2F3, MYC*	DNA damage, proliferation	[76-78]
miR-31	*FZD3, ITGA5, M-RIP, MMP16, RDX, RHOA*	Metastasis	[61]
Oncogenic miRNAs			
miR-21	*BCL-2, TPM1, PDCD4, PTEN, MASPIN*	Apoptosis	[64-67]
miR-155	*RHOA*	TGF-β signaling	[68]
miR-10b	*HOXD10*	Metastasis	[69]
miR-373/520c	*CD44*	Metastasis	[79]

miRNAs and their targets in breast cancer. Several studies have identified altered miRNA expression profiles in breast cancer. The loss of several tumor suppressor miRNAs and overexpression of oncogenic miRNAs leads to loss of regulation of several cellular functions that may be involved in breast cancer pathogenesis.

breast cancer pathogenesis. The tumor suppressive role for miR-206 in breast cancer was further substantiated by the recent demonstration that miR-206, miR-335 and miR-126 are potently downregulated in metastatic breast cancer cells compared to parental cells [35]. Importantly, restoring the expression of these three miRNAs reduced their invasive capacity. The authors identified four direct targets of miR-335, *PTPRN2*, *MERTK*, *TNC* and *SOX4*; among them, *TNC* (encoding an extracellular matrix component) and *SOX4* (encoding a transcription factor that is involved in tumorigenesis) were shown to be functional targets implicated in metastasis. Restoring miR-206 expression in metastatic cells did not influence their proliferation or sensitivity to apoptosis, but altered cellular morphology, possibly contributing to a decrease in cell motility that could limit the migration of metastatic cells [35]. These findings suggest that miR-206 could be a novel candidate for breast cancer therapy.

miR-17-5p

miR-17-5p, also known as miR-91, is located on chromosome 13q31, a genomic region that undergoes loss of heterozygosity in multiple cancers, including breast cancer [36]. The oncogene *AIB1* (amplified in breast cancer) is a direct target of miR-17-5p [37]. The protein encoded by the *AIB1* gene is a steroid receptor co-activator that enhances the transcriptional activity of *ERα*, *E2F1* (which is also directly regulated by miR-17-5p) and other transcription factor genes. miR-17-5p represses the translation of *AIB1* mRNA, thereby inhibiting the function of E2F1 and ERα. Downregulation of AIB1 by miR-17-5p results in the suppression of estrogen-stimulated proliferation and estrogen/ER-independent breast cancer cell proliferation [37]. In breast cancer cells, the gene cyclin D1 (*CCND1*), which is overexpressed in approximately 50% of human cancers, was recently identified as a direct target of miR-17-5p [38]. miR-17-5p inhibits the proliferation of breast cancer cells by suppressing cyclin D1 protein synthesis and this effect was abrogated by a *CCND1* siRNA in cyclin D1-deficient breast cancer cells. This study also identified a regulatory mechanism in which cyclin D1 induces miR-17-5p expression, thus providing a negative feedback loop that limits cyclin D1 expression.

miR-125a and miR-125b

The *HER2* (*ERBB2*) gene encodes a member of the epidermal growth factor receptor family of tyrosine kinases. This protein has no ligand-binding domain of its own and cannot, therefore, bind growth factors. However, it binds tightly to other ligand-bound epidermal growth factor receptor family members to form a heterodimer, stabilizing ligand-binding activity and enhancing kinase-mediated activation of downstream signaling pathways, such as those involving mitogen-activated protein kinase (MAPK) and phosphatidylinositol-3 kinase [39]. Amplification and/or overexpression of this gene has been reported in numerous cancers, including breast tumors. miR-125a and miR-125b are downregulated in HER2-amplified and HER2-overexpressing breast cancers [40]. These two miRNAs are potential tumor suppressors and their overexpression in SKBR3 cells (a HER2-dependent human breast cancer cell line) suppresses *HER2* and *HER3* mRNA and protein levels, leading to a reduction in anchorage-dependent growth, cell motility, and invasiveness [41]. However, this influence was subtle in non-transformed HER2-independent breast cancer cells (MCF10A).

miR-200 family

Invasion and metastasis are the hallmarks of malignant tumor progression. Increasing evidence indicates that activation of the embryonic program 'epithelial-mesenchymal transition' (EMT) promotes these processes by allowing detachment of cells from each other, thereby increasing tumor cell mobility and dissemination [42]. In recent years, it has become evident that aberrant activation of EMT is responsible for the malignant transformation of many human cancers. EMT is activated by EMT-inducing transcriptional repressors, including members of the Snail family and the ZFH family of transcription factors [43]. These proteins inhibit the transcription of epithelial genes, such as that encoding E-cadherin. Recently, ZEB1, a ZFH-family member, was shown to be a crucial EMT activator in human cancers, including breast cancer [44,45], and to promote metastasis of tumor cells in a mouse xenograft model [46]. The miR-200 family of miRNAs has been shown to be crucial inducers of an epithelial phenotype by suppressing the expression of the EMT inducers ZEB1 and ZEB2. Gregory *et al.* [47] induced EMT in MDCK cells (human Madin Darby canine kidney epithelial cells) either by transforming growth factor (TGF)-β or by stable expression of the tyrosine phosphatase Pez. The authors found that all five members of the miR-200 family, which are clustered in two genomic locations, were downregulated upon TGF-β-induced EMT. Functional studies showed that inhibition of miR-200a, b, and c in MDCK cells triggered TGF-β-independent EMT that was associated with increased expression of vimentin, fibronectin and N-cadherin, and downregulation of E-cadherin. Conversely, overexpression of miR-200 family members in mesenchymal cells initiated mesenchymal-to-epithelial transition. Luciferase reporter assays under the control of *ZEB1* and *ZEB2* 3′ UTR verified that these genes are direct targets of the miR-200 family. The clinical relevance of these findings was further suggested by additional analyses of human breast cancer tissues showing that a decrease in miRNA-200 family miRNAs was associated with highly aggressive, metaplastic breast tumors. Their finding that *ZEB1* and *ZEB2* expression is controlled by the miR-200 family suggests that down-regulation of these miRNAs is an essential early step in tumor metastasis.

Similar overlapping results were obtained in another study, which employed the NCI-60 panel of cell lines, representing 60 cell lines of different human tumor types [48]. By using two markers, the authors divided the panel into two major groups, characterized either by an epithelial phenotype (marked by E-cadherin) or a mesenchymal phenotype (marked by vimentin) and found that the miR-200 family was strongly associated with the epithelial phenotype. Overexpression of miR-200 induced epithelial differentiation in undifferentiated breast cancer cells (MDA-MB231) whereas antagonizing it induced an EMT phenotype in the colorectal cancer cell line HCT116. *ZEB1* and *ZEB2* were again detected as targets of the miR-200 family members. These and other recent reports on the miR-200 family [49,50] add important miRNAs to the growing list of tumor-associated miRNAs. Furthermore, a recent study showed that the expression of the miR-200 family was decreased by Akt2, suggesting that, in many cases, breast cancer metastasis may be under the control of the Akt-miR-200-E-cadherin pathway [51]. Interestingly, a recent study showed that expression of miR-200 unexpectedly enhanced macroscopic metastases in mouse breast cancer cell lines [52]. Their results suggest that, for some tumors, tumor colonization at metastatic sites might be enhanced by mesenchymal-to-epithelial transition and that the epithelial nature of a tumor may not predict metastatic outcome. Although this paper is the first to show the direct enhancement of metastasis by the miR-200 family, changes in miR-200 family levels have been associated with enhanced tumorigenesis. For instance, the miR-200 family is upregulated in human ovarian cancers compared to normal ovarian tissue [29]. Moreover, overexpression of the miR-200 family significantly correlates with decreased survival. Patients with ovarian tumors with high miR-200a expression have been shown to have approximately 50% decrease in median survival time compared to those lacking significant miR-200a expression. Furthermore, the region on chromosome 1 that encodes the miR-200 cluster was also found to be amplified in several epithelial cancers - ovarian cancer, breast cancer and melanoma [53]. A new study [54] revealed the roles of three miRNA clusters (miR-200c-141, miR-200b-200a-429 and miR-183-96-182) in the regulation of self-renewal in cancer stem cells and normal stem cells. These miRNAs were downregulated in human breast cancer stem cells, normal human and mouse mammary stem/progenitor cells as well as embryonal carcinoma cells. The authors provided evidence that *BMI1* (a gene that promotes stem cell self-renewal) is specifically inhibited by these miRNAs and showed that ectopic overexpression of miR-200c in embryonal carcinoma cells resulted in growth retardation and neural differentiation and suppressed tumorigenicity of breast cancer stem cells *in vivo* [55]. Taken together, these studies suggest that miRNAs of the miR-200 family play important roles in regulating tumor progression and metastasis.

let-7 family

let-7, one of the founding members of the miRNA family, was originally discovered to play a role in the developmental timing of *C. elegans* [56] and is conserved throughout the animal phyla. let-7 is poorly expressed or

deleted in many human cancers. Recent data from both hematologic malignancies and solid tumors suggest that each includes minor populations of cells that are capable of tumor initiation [57]. These tumor-initiating cells (T-ICs) have properties of tumor stem cells and can undergo asymmetric cell division and self renewal. T-ICs comprise only a minor fraction among the bulk of the more differentiated cells in the tumor. Based on the cancer stem cell hypothesis, T-ICs are responsible for the initiation, progression, metastasis and resistance to therapy. Breast T-ICs (BT-ICs) can be enriched by purifying $CD44+CD24^{-/low}$ cells by sorting or by purifying spherical clusters of self-replicating cells (mammospheres) from cell suspensions.

A recent study compared miRNA expression in self-renewing and differentiated cells from breast cancer lines and found that the expression of let-7 was strongly reduced in BT-ICs and increased with differentiation [58]. Introducing let-7 in BT-ICs reduced their proliferative capacity, their ability to form mammospheres and tumor formation and metastasis *in vivo*. Conversely, knocking down let-7 enhanced self-renewal of non-T-ICs *in vitro*. Known oncogenic targets of let-7, such as *H-RAS* and *HMGA2*, were downregulated by let-7 overexpression. Silencing *H-RAS* in a BT-IC-enriched cell line reduced self-renewal but had no effect on differentiation, while knocking down *HMGA2* enhanced differentiation but did not affect self-renewal. Their results suggest that let-7 regulates multiple BT-IC stem cell-like properties and that let-7 may offer a unique opportunity to attack tumor stem cells using therapeutic RNA. Delivery of the *let-7* miRNA to tumors could potentially deplete stem cells by inducing cellular differentiation.

A recent report linked the protein RKIP (Raf kinase inhibitory protein) to let-7 and breast cancer metastasis. RKIP (also called PEBP1) inhibits MAPK, G protein-coupled receptor kinase-2, and NF-κB signaling cascades [59]. The authors found that RKIP inhibited invasion by metastatic breast cancer cells and repressed breast tumor cell intravasation and bone metastasis in a mouse model. Inhibition of MAPK decreased transcription of *LIN28* by MYC; in turn, downregulation of LIN28, an inhibitor of let-7 biogenesis, enhanced let-7 expression in breast cancer cells leading to decreased expression of HMGA2, a chromatin-remodeling protein that activates pro-invasive and prometastatic genes, including Snail. Collectively, their findings indicate that RKIP suppresses invasion and metastasis via a signaling cascade that involves MAPK, MYC, LIN28, let-7, and downstream let-7 targets such as HMGA2.

miR-34a

miR-34a is one of several miRNAs that are downregulated in multiple cancers [26] and has been shown to be transcriptionally regulated by p53. In the context of breast cancer, only one study [60] has shown that miR-34a levels were lower in triple negative and mesenchymal breast cancer cell lines compared with normal epithelial lines and HER-2+ lines. The authors propose that p53 mutations in these subtypes of breast cancer might contribute to low miR-34a expression. To elucidate the relationship between miR-34a and radiosensitivity in these cells, the authors compared the sensitivity of a normal breast epithelial line (HMEC), a HER-2+ line (UACC812) and a mesenchymal line (MDA-MB-231) and found that the MDA-MB-231 cell line (with low miR-34a levels) was significantly more sensitive to radiation than the HMECs or UACC-812 (with high miR-34a levels). Increasing the levels of miR-34a protected the MDA-MB-231 cells from radiation-induced cell death, and downregulating it had the converse effect. These results show that miR-34a is necessary for the survival of MDA-MB-231 cells from non-apoptotic cell death and suggest that miR-34a may have therapeutic potential in breast cancers, since antagonizing miR-34a increases the sensitivity of breast cancer cells towards radiation.

miR-31

miR-31 was recently shown to prevent metastasis at multiple steps by inhibiting the expression of prometastatic genes [61]. miR-31 is expressed in normal breast cells and its abundance was shown to be dependent on the metastatic state of the tumor. It is moderately decreased in non-metastatic breast cancer cell lines and is almost undetectable in metastatic mouse and human breast cancer cell lines. Importantly, the authors demonstrated that introducing miR-31 in metastatic breast cancer cells suppressed metastasis-related functions (motility, invasion and resistance to anoikis) *in vitro* and metastasis *in vivo*. Notably, they found that although miR-31 over-expressing breast cancer cells formed larger and more proliferative tumors, these tumors were well encapsulated and thereby less invasive, suggesting that miR-31 inhibits metastasis in the early stages of the metastasis cascade. Injecting miR-31 overexpressing cells directly into the circulation impeded the ability of the cells to survive and form secondary tumors in the lung, suggesting that miR-31 inhibits metastasis at multiple steps of the metastatic cascade. Conversely, inhibition of miR-31 function increased invasiveness and promoted metastasis *in vivo*. To identify functionally relevant mRNA targets of miR-31 through which it might exert its antimetastatic effects, the authors employed Gene Ontology analysis on the putative targets of miR-31 predicted by the miRNA-target prediction algorithms TargetScan and PicTar. Using this approach, they validated six targets of miR-31 in human breast cancer cells: frizzled3 (*Fzd3*), integrin α-5 (*ITGA5*), myosin

phosphatase-Rho-interacting protein (*M-RIP*), matrix metallopeptidase 16 (*MMP16*), radixin (*RDX*) and the ras homolog gene family member A (*RhoA*). Interestingly, they showed that re-expression of three of these genes, *ITGA5*, *RDX* and *RhoA*, in metastatic breast cancer cells abrogated the motility defects and reversed the impaired invasion and anoikis resistance defects conferred by miR-31, suggesting that these three genes are functionally important targets of miR-31. Taken together, their findings demonstrate that miR-31 may also be an attractive therapeutic target for breast cancer as it exerts its antimetastatic effect by targeting multiple pro-metastatic genes of the metastasis cascade.

Oncogenic miRNAs

The function of four well-established oncogenic and/or metastasis promoting miRNAs (Table 1) are discussed below.

miR-21

miR-21 has been found to be consistently overexpressed in many tumors, including breast cancer [28,62]. Consistent with these findings, miR-21 was also shown to be highly upregulated in breast tumors compared to the matched normal breast tissues among 157 human miRNAs analyzed by real-time RT-PCR arrays [63], suggesting that miR-21 may function as an oncogene. The latter study also examined the oncogenic role of miR-21 in cell culture as well as in mouse xenografts. Although the influence of knocking down miR-21 on cell survival was marginal (25% decrease), the authors found that this growth inhibition increased when the transfected MCF-7 cells were treated with the anticancer drug topotecan, suggesting that suppression of miR-21 can sensitize tumor cells to anticancer agents. To further investigate whether suppression of miR-21 alone influences tumorigenesis, they antagonized miR-21 in MCF-7 cells and then injected them into mammary pads of female nude mice. Tumors derived from MCF-7 cells transfected with anti-miR-21 were one-half the size of those derived from the cells transfected with the negative control, while anti-Ki-67 indicated that the reduced tumor growth was likely due to a lower proliferation caused by anti-miR-21. These results show that miR-21 is an oncogenic miRNA that plays an important role in tumorigenesis. The authors also found that antagonizing miR-21 caused apoptosis in MCF-7 cells, which was associated with lower expression of Bcl-2 protein in the anti-miR-21-transfected MCF-7 cells and also in tumors derived from these cells, suggesting that Bcl-2 may be an indirect target of miR-21. To further elucidate the molecular mechanisms that regulate the function of miR-21, proteomic analysis of the above-mentioned xenograft tumors revealed that the tumor suppressor protein tropomyosin 1 (TPM1), which is known to be downregulated in breast cancer epithelial cell lines, was a target of miR-21 [64]. In the context of breast cancer, miR-21 has also been shown to inhibit the expression of the tumor suppressor PDCD4 (programmed cell death-4). Although depletion of PDCD4 protein in MCF-7 cells by *PDCD4* siRNA had no effect on cellular proliferation, it significantly alleviated the anti-proliferative effect of miR-21 inhibition, underscoring an essential role for PDCD4 as a mediator of the biological effects of miR-21 in breast cancer cells [65]. Another important target of miR-21 is the tumor suppressor gene phosphatase and tensin homolog (*PTEN*) [66]. In addition to *TPM1* and *PDCD4*, one study also identified *Maspin* as a direct target of miR-21 [67]. These three genes were shown to reduce invasiveness of a metastatic breast cancer cell line [64]. Their findings further establish miR-21 as an oncogenic miRNA and suggest that miR-21 has a role not only in tumor growth but also in invasion and tumor metastasis by targeting multiple anti-metastatic genes.

miR-155

miR-155 is over-expressed in a number of human malignancies, including breast cancer [28,62]. A recent study has shown that miR-155 is upregulated in normal mouse mammary gland epithelial cells (NMuMG cells) by the TGF-β/Smad4 pathway and mediates TGF-β-induced EMT and cell invasion [68]. To investigate the role of miR-155 in TGF-β-induced EMT, the authors found that ectopic expression of miR-155 in NMuMG cells disrupted proper tight junction formation and promoted cell migration and invasion. Conversely, antagonizing miR-155 in NMuMG cells reduced the occurrence of TGF-β-induced EMT and cell migration and invasion. miR-155 directly inhibited the expression of *RhoA*, a gene that regulates many cellular processes, including cell adhesion, motility, and polarity, and is an important modulator of cell junction formation and stability. Importantly, miR-155-induced phenotypes were restored by expressing a miR-155 insensitive version of *RhoA* in miR-155 overexpressing cells. These findings suggest that miR-155 is regulated by the TGF-β/Smad4 pathway and downregulates RhoA protein expression to drive EMT progression. They further demonstrated that miR-155 is associated with cancer invasiveness in human primary breast carcinoma by showing that miR-155 is highly expressed in invasive tumors but not in noninvasive cancer tissues. These results implicate miR-155 in EMT and invasion as observed in NMuMG cells and suggest that miR-155 may play a critical role in breast cancer metastasis.

miR-10b

miR-10b was the first miRNA found to influence the metastatic potential of human cancer cells [69]. Unlike

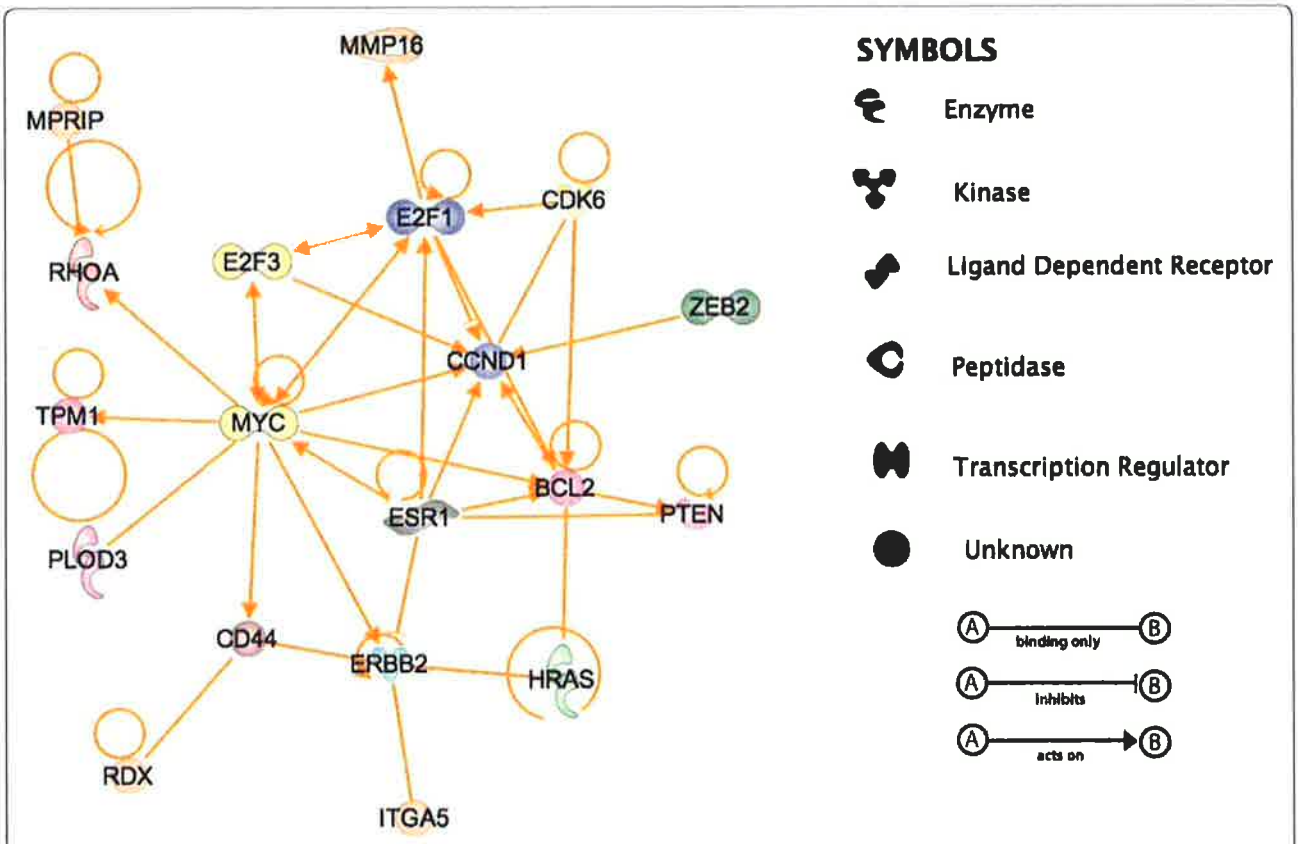


Figure 1. Gene interaction network analysis of miRNA targets in breast cancer. Determined using Ingenuity software, the direct interaction network of 34 published targets of 11 miRNAs implicated in breast cancer pathogenesis is shown: miR-206 (grey), miR-17-5p (blue), miR-125a/b (cyan), miR-200 (dark green), let-7 (green), miR-34 (yellow), miR-31 (orange), miR-21 (pink), miR-155 (red), and miR-373/520c (maroon). The targets form a highly connected network centered on MYC. Arrows indicate protein-protein or protein-DNA interactions, suggesting these targets coordinate the expression and/or function of one another.

miR-155, which is overexpressed in many breast tumors, miR-10b was highly expressed only in metastatic cancer cells and was found to promote cell migration and invasion *in vitro* and initiate tumor invasion and metastasis *in vivo*. Expression of miR-10b was found to be induced by the transcription factor Twist; in turn, miR-10b inhibits the translation of the transcription factor homeobox D10 (HOXD10), resulting in a cascade of cellular alterations that include expression of the prometastatic gene *RHOC* (ras homologue gene family member C), a gene that promotes cancer cell migration and invasion.

miR-373/520c family

miRNAs of the miR-373/520c family were identified [60] as prometastatic using a forward genetic screen involving overexpression of almost 450 miRNAs in a non-metastatic human breast cancer cell line (MCF-7 cells). MCF-7 cells were transduced with these miRNAs and subjected to a trans-well cell migration assay to identify miRNAs that stimulate cell migration. The authors found that miR-373 and miR-520c promoted cancer cell migration and invasion *in vitro* and *in vivo*. They also showed that some cancer cell lines required miR-373 for migration and that these miRNAs elicited a migratory phenotype by inhibiting expression of the metastasis repressor CD44. Importantly, ectopic overexpression of CD44 without its 3′ UTR significantly reduced the migration of MCF-7 cells that express miR-373 or miR-520c, suggesting that downregulation of CD44 is required for the migration phenotype of MCF-7 cells expressing miR-373 and miR-520c. Finally, the authors showed that miR-373 was upregulated whereas CD44 was decreased in clinical breast cancer metastasis samples, further implicating its role in breast cancer metastasis.

Gene interaction network for breast cancer miRNA targets

The above-mentioned investigations on miRNAs implicated in breast cancers substantially advance our knowledge about the roles of specific miRNAs in breast cancer development and metastasis. It is clear that each of these miRNAs inhibit the expression of many genes,

suggesting that comprehensive regulation can be achieved by antagonizing or over-expressing a single miRNA. Moreover, concomitant deregulation of these miRNAs would consequently alter the expression of many genes, thereby inducing tumorigenesis.

To better understand how the functions of these experimentally validated breast cancer miRNAs and their gene targets might be integrated within the pathogenesis of breast cancer, we have performed a gene interaction network analysis (Figure 1) similar to our previous study [70]. To do this, we generated a list of 34 genes known to be altered by the 11 miRNAs discussed above. Among these genes, 19 formed a well-connected gene interaction network in which MYC, a target of miR-34a, was the central node (10 connections). Other highly interacting genes were *BCL-2*, *E2F1*, *CCND1* and *ESR1*, with seven, seven, seven and six interactions, respectively.

These results suggest that breast cancer is associated with changes in the expression of multiple miRNAs that, in turn, disrupt a network of genes that either activate or inhibit each other's transcription or interact directly via protein-protein interactions. For instance, downregulation of miR-206 would enhance *ESR1* expression, consequently increasing *MYC* transcription. MYC protein will be further elevated since loss of miR-34a would enhance its translation. Since CCND1, E2F1 and E2F3 are known to be activated by MYC, increased MYC would in turn enhance their expression. Moreover, down-modulation of miR-17-5p (regulates *CCND1* and *E2F1*) and miR-34a (regulates *E2F3*, *CCND1* and *CDK6*) would further elevate levels of these proteins. Our findings suggest that decreased expression of multiple tumor suppressor miRNAs leads to enhanced transcription and translation of these oncogenes in breast cancer via direct or indirect mechanisms.

The role of oncogenic or metastasis-promoting miRNAs is also important. For example, increased levels of miR-21, miR-31 and miR-373/520c would ensure that the protein synthesis of TPM1 (regulated by miR-21), CD44 (regulated by miR-373/520c), and MMP16 (regulated by miR-31) is repressed in breast cancer cells. In most cases, the protein encoded by a gene is up- or downregulated due to increased/decreased transcription and translation. Thus, tumor suppressor or oncogenic miRNAs regulate the transcription of some of these genes by targeting transcription factors and also provide an additional layer of regulation by regulating their translation. Although many of these gene regulatory events will likely occur at different stages of tumorigenesis and metastasis, it is clear that loss of miRNA regulation leads to a cascade of events that alter gene interaction networks important for breast cancer progression.

In addition to the above mentioned miRNAs, several other miRNAs, including miR-7, miR-128a, miR-210, miR-27b, miR-335, miR-126, miR-145 and miR-27a, have also been implicated in breast cancer. Uncovering the genes under the control of these miRNAs and how they integrate into the breast cancer gene interaction network will further aid our understanding of the disease. Furthermore, miRNA profiling is emerging as a powerful diagnostic tool to characterize features of different tumor types. This has been particularly useful in breast cancer, as miRNA signatures can unequivocally distinguish normal and malignant breast tissue and discriminate between breast cancer subtypes [71].

Conclusion

The usefulness of miRNA-based breast cancer therapy has been explored by these emerging studies that highlight their importance in breast cancer. A better understanding of the network of genes and cellular pathways regulated by these miRNAs will undoubtedly enable us to understand breast cancer pathogenesis and therapy. To accomplish this, identifying the genome-wide targets of these miRNAs is essential. Moreover, it is yet to be seen how these breast cancer-associated miRNAs will move from the laboratory bench into a clinical setting for the treatment of breast cancer. In this regard, the delivery of miRNA inhibitors or miRNA mimics specifically to tumor cells is one of the greatest challenges. Viral vector systems seem to be a reasonable option for efficient and organ-specific delivery [72], and it is likely that miRNA-based therapeutics will become a reality in the near future.

Abbreviations

BT-IC = breast tumor-initiating cell; EMT = epithelial-mesenchymal transition; ER = estrogen receptor; MAPK = mitogen-activated protein kinase; MDCK = Madin Darby canine kidney; miRNA = microRNA; siRNA = small interfering RNA; TGF = transforming growth factor; T-IC = tumor-initiating cell; UTR = untranslated region.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Myriam Gorospe (NIH) for careful reading of the article. Due to space constraints this is not an exhaustive review of the literature on the subject and we apologize for not citing all publications.

Published: 19 March 2010

References

1. Borchert GM, Lanier W, Davidson BL: **RNA polymerase III transcribes human microRNAs.** *Nat Struct Mol Biol* 2006, **13**:1097-1101.
2. Tomari Y, Zamore PD: **Perspective: machines for RNAi.** *Genes Dev* 2005, **19**:517-529.
3. Filipowicz W, Bhattacharyya SN, Sonenberg N: **Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?** *Nat Rev Genet* 2008, **9**:102-114.
4. Lehmann U, Hasemeier B, Christgen M, Müller M, Römermann D, Länger F, Kreipe H: **Epigenetic inactivation of microRNA gene hsa-mir-9-1 in human breast cancer.** *J Pathol* 2008, **214**:17-24.
5. Lal A, Kim HH, Abdelmohsen K, Kuwano Y, Pullmann R Jr, Srikantan S, Subrahmanyam R, Martindale JL, Yang X, Ahmed F, Navarro F, Dykxhoorn D, Lieberman J, Gorospe M: **p16(INK4a) translation suppressed by miR-24.**

PLoS ONE 2008, **3**:e1864.

6. Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos I: **MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation.** *Nature* 2008, **455**:1124-1128.
7. Abdelmohsen K, Srikantan S, Kuwano Y, Gorospe M: **miR-519 reduces cell proliferation by lowering RNA-binding protein HuR levels.** *Proc Natl Acad Sci USA* 2008, **105**:20297-20302.
8. Forman JJ, Legesse-Miller A, Coller HA: **A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence.** *Proc Natl Acad Sci USA* 2008, **105**:14879-14884.
9. Lytle JR, Yario TA, Steitz JA: **Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5'UTR as in the 3'UTR.** *Proc Natl Acad Sci USA* 2007, **104**:9667-9672.
10. Olsen PH, Ambros V: **The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation.** *Dev Biol* 1999, **216**:671-680.
11. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output.** *Nature* 2008, **455**:64-71.
12. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455**:58-63.
13. Vasudevan S, Tong Y, Steitz JA: **Switching from repression to activation: microRNAs can up-regulate translation.** *Science* 2007, **318**:1931-1934.
14. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136**:215-233.
15. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120**:15-20.
16. Zhao Y, Srivastava D: **A developmental view of microRNA function.** *Trends Biochem Sci* 2007, **32**:189-197.
17. Ventura A, Jacks T: **MicroRNAs and cancer: short RNAs go a long way.** *Cell* 2009, **136**:586-591.
18. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer.** *Nat Rev Cancer* 2006, **6**:259-269.
19. Croce CM: **Causes and consequences of microRNA dysregulation in cancer.** *Nat Rev Genet* 2009, **10**:704-714.
20. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM: **bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*.** *Cell* 2003, **113**:25-36.
21. Lee RC, Feinbaum RL, Ambros V: **The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75**:843-854.
22. Calin GA, Liu CG, Sevignani C, Ferracin M, Felli N, Dumitru CD, Shimizu M, Cimmino A, Zupo S, Dono M, Dell'Aquila ML, Alder H, Rassenti L, Kipps TJ, Bullrich F, Negrini M, Croce CM: **MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias.** *Proc Natl Acad Sci USA* 2004, **101**:11755-11760.
23. Calin GA, Croce CM: **MicroRNA signatures in human cancers.** *Nat Rev Cancer* 2006, **6**:857-866.
24. Jiang J, Lee EJ, Gusev Y, Schmittgen TD: **Real-time expression profiling of microRNA precursors in human cancer cell lines.** *Nucleic Acids Res* 2005, **33**:5394-5403.
25. Murakami Y, Yasuda T, Saigo K, Urashima T, Toyoda H, Okanoue T, Shimotohno K: **Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues.** *Oncogene* 2006, **25**:2537-2545.
26. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR, Israel MA: **Characterization of microRNA expression levels and their biological correlates in human cancer cell lines.** *Cancer Res* 2007, **67**:2456-2468.
27. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers.** *Nature* 2005, **435**:834-838.
28. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Ménard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM: **MicroRNA gene expression deregulation in human breast cancer.** *Cancer Res* 2005, **65**:7065-7070.
29. Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, Taccioli C, Volinia S, Liu CG, Alder H, Calin GA, Ménard S, Croce CM: **MicroRNA signatures in human ovarian cancer.** *Cancer Res* 2007, **67**:8699-8707.
30. Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, Iorio MV, Visone R, Sever NI, Fabbri M, Iuliano R, Palumbo T, Pichiorri F, Roldo C, Garzon R, Sevignani C, Rassenti L, Alder H, Volinia S, Liu CG, Kipps TJ, Negrini M, Croce CM: **A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia.** *N Engl J Med* 2005, **353**:1793-1801.
31. Schetter AJ, Leung SY, Sohn JJ, Zanetti KA, Bowman ED, Yanaihara N, Yuen ST, Chan TL, Kwong DL, Au GK, Liu CG, Calin GA, Croce CM, Harris CC: **MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma.** *JAMA* 2008, **299**:425-436.
32. Kumar MS, Lu J, Mercer KL, Golub TR, Jacks T: **Impaired microRNA processing enhances cellular transformation and tumorigenesis.** *Nat Genet* 2007, **39**:673-677.
33. Adams BD, Furneaux H, White BA: **The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor-alpha (ERalpha) and represses ERalpha messenger RNA and protein expression in breast cancer cell lines.** *Mol Endocrinol* 2007, **21**:1132-1147.
34. Leivonen SK, Mäkelä R, Ostling P, Kohonen P, Haapa-Paananen S, Kleivi K, Enerly E, Aakula A, Hellström K, Sahlberg N, Kristensen VN, Børresen-Dale AL, Saviranta P, Perälä M, Kallioniemi O: **Protein lysate microarray analysis to identify microRNAs regulating estrogen receptor signaling in breast cancer cell lines.** *Oncogene* 2009, **28**:3926-3936.
35. Tavazoie SF, Alarcón C, Oskarsson T, Padua D, Wang Q, Bos PD, Gerald WL, Massagué J: **Endogenous human microRNAs that suppress breast cancer metastasis.** *Nature* 2008, **451**:147-152.
36. Eiriksdottir G, Johannesdottir G, Ingvarsson S, Björnsdottir IB, Jonasson JG, Agnarsson BA, Hallgrimsson J, Gudmundsson J, Egilsson V, Sigurdsson H, Barkardottir RB: **Mapping loss of heterozygosity at chromosome 13q: loss at 13q12-q13 is associated with breast tumour progression and poor prognosis.** *Eur J Cancer* 1998, **34**:2076-2081.
37. Hossain A, Kuo MT, Saunders GF: **Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA.** *Mol Cell Biol* 2006, **26**:8191-8201.
38. Yu Z, Wang C, Wang M, Li Z, Casimiro MC, Liu M, Wu K, Whittle J, Ju X, Hyslop T, McCue P, Pestell RG: **A cyclin D1/microRNA 17/20 regulatory feedback loop in control of breast cancer cell proliferation.** *J Cell Biol* 2008, **182**:509-517.
39. Baselga J, Swain SM: **Novel anticancer targets: revisiting ERBB2 and discovering ERBB3.** *Nat Rev Cancer* 2009, **9**:463-475.
40. Mattie MD, Benz CC, Bowers J, Sensinger K, Wong L, Scott GK, Fedele V, Ginzinger D, Getts R, Haqq C: **Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies.** *Mol Cancer* 2006, **5**:24.
41. Scott GK, Goga A, Bhaumik D, Berger CE, Sullivan CS, Benz CC: **Coordinate suppression of ERBB2 and ERBB3 by enforced expression of micro-RNA miR-125a or miR-125b.** *J Biol Chem* 2007, **282**:1479-1486.
42. Thiery JP, Acloque H, Huang RYJ, Nieto MA: **Epithelial-mesenchymal transitions in development and disease.** *Cell* 2009, **139**:871-890.
43. Peinado H, Olmeda D, Cano A: **Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype?** *Nat Rev Cancer* 2007, **7**:415-428.
44. Aigner K, Dampier B, Descovich L, Mikula M, Sultan A, Schreiber M, Mikulits W, Brabletz T, Strand D, Obrist P, Sommergruber W, Schweifer N, Wernitznig A, Beug H, Foisner R, Eger A: **The transcription factor ZEB1 (deltaEF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity.** *Oncogene* 2007, **26**:6979-6988.
45. Dohadwala M, Yang SC, Luo J, Sharma S, Batra RK, Huang M, Lin Y, Goodglick L, Krysan K, Fishbein MC, Hong L, Lai C, Cameron RB, Gemmill RM, Drabkin HA, Dubinett SM: **Cyclooxygenase-2-dependent regulation of E-cadherin: prostaglandin E(2) induces transcriptional repressors ZEB1 and snail in non-small cell lung cancer.** *Cancer Res* 2006, **66**:5338-5345.
46. Spaderna S, Schmalhofer O, Wahlbuhl M, Dimmler A, Bauer K, Sultan A, Hlubek F, Jung A, Strand D, Eger A, Kirchner T, Behrens J, Brabletz T: **The transcriptional repressor ZEB1 promotes metastasis and loss of cell polarity in cancer.** *Cancer Res* 2008, **68**:537-544.
47. Gregory PA, Bert AG, Paterson EL, Barry SC, Tsykin A, Farshid G, Vadas MA, Khew-Goodall Y, Goodall GJ: **The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1.** *Nat Cell Biol* 2008, **10**:593-601.
48. Park S-M, Gaur AB, Lengyel E, Peter ME: **The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2.** *Genes Dev* 2008, **22**:894-907.

49. Hurteau GJ, Carlson JA, Roos E, Brock GJ: **Stable expression of miR-200c alone is sufficient to regulate TCF8 (ZEB1) and restore E-cadherin expression.** *Cell Cycle* 2009, **8**:2064-2069.
50. Korpal M, Lee ES, Hu G, Kang Y: **The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2.** *J Biol Chem* 2008, **283**:14910-14914.
51. Iliopoulos D, Polytarchou C, Hatziapostolou M, Kottakis F, Maroulakou IG, Struhl K, Tsichlis PN: **MicroRNAs differentially regulated by Akt isoforms control EMT and stem cell renewal in cancer cells.** *Sci Signal* 2009, **2**:ra62.
52. Dykxhoorn DM, Wu Y, Xie H, Yu F, Lal A, Petrocca F, Martinvalet D, Song E, Lim B, Lieberman J: **miR-200 enhances mouse breast cancer cell colonization to form distant metastases.** *PLoS ONE* 2009, **4**:e7181.
53. Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A, Liang S, Naylor TL, Barchetti A, Ward MR, Yao G, Medina A, O'brien-Jenkins A, Katsaros D, Hatzigeorgiou A, Gimotty PA, Weber BL, Coukos G: **microRNAs exhibit high frequency genomic alterations in human cancer.** *Proc Natl Acad Sci USA* 2006, **103**:9136-9141.
54. Shimono Y, Zabala M, Cho RW, Lobo N, Dalerba P, Qian D, Diehn M, Liu H, Panula SP, Chiao E, Dirbas FM, Somlo G, Pera RA, Lao K, Clarke MF: **Downregulation of miRNA-200c links breast cancer stem cells with normal stem cells.** *Cell* 2009, **138**:592-603.
55. Wellner U, Schubert J, Burk UC, Schmalhofer O, Zhu F, Sonntag A, Waldvogel B, Vannier C, Darling D, zur Hausen A, Brunton VG, Morton J, Sansom O, Schüler J, Stemmler MP, Herzberger C, Hopt U, Keck T, Brabletz S, Brabletz T: **The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs.** *Nat Cell Biol* 2009, **11**:1487-1495.
56. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*.** *Nature* 2000, **403**:901-906.
57. Clarke MF, Fuller M: **Stem cells and cancer: two faces of eve.** *Cell* 2006, **124**:1111-1115.
58. Yu F, Yao H, Zhu P, Zhang X, Pan Q, Gong C, Huang Y, Hu X, Su F, Lieberman J, Song E: **let-7 regulates self renewal and tumorigenicity of breast cancer cells.** *Cell* 2007, **131**:1109-1123.
59. Dangi-Garimella S, Yun J, Eves EM, Newman M, Erkeland SJ, Hammond SM, Minn AJ, Rosner MR: **Raf kinase inhibitory protein suppresses a metastasis signalling cascade involving LIN28 and let-7.** *EMBO J* 2009, **28**:347-358.
60. Kato M, Paranjape T, Müller RU, Ullrich R, Nallur S, Gillespie E, Keane K, Esquela-Kerscher A, Weidhaas JB, Slack FJ: **The mir-34 microRNA is required for the DNA damage response in vivo in *C. elegans* and *in vitro* in human breast cancer cells.** *Oncogene* 2009, **28**:2419-2424.
61. Valastyan S, Reinhardt F, Benaich N, Calogrias D, Szász AM, Wang ZC, Brock JE, Richardson AL, Weinberg RA: **A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis.** *Cell* 2009, **137**:1032-1046.
62. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: **A microRNA expression signature of human solid tumors defines cancer gene targets.** *Proc Natl Acad Sci USA* 2006, **103**:2257-2261.
63. Si M-L, Zhu S, Wu H, Lu Z, Wu F, Mo Y-Y: **miR-21-mediated tumor growth.** *Oncogene* 2007, **26**:2799-2803.
64. Zhu S, Si M-L, Wu H, Mo Y-Y: **MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1).** *J Biol Chem* 2007, **282**:14328-14336.
65. Frankel LB, Christoffersen NR, Jacobsen A, Lindow M, Krogh A, Lund AH: **Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells.** *J Biol Chem* 2008, **283**:1026-1033.
66. Qi L, Bart J, Tan LP, Platteel I, Sluis Tvd, Huitema S, Harms G, Fu L, Hollema H, Berg Avd: **Expression of miR-21 and its targets (PTEN, PDCD4, TM1) in flat epithelial atypia of the breast in relation to ductal carcinoma *in situ* and invasive carcinoma.** *BMC Cancer* 2009, **9**:163.
67. Zhu S, Wu H, Wu F, Nie D, Sheng S, Mo Y-Y: **MicroRNA-21 targets tumor suppressor genes in invasion and metastasis.** *Cell Res* 2008, **18**:350-359.
68. Kong W, Yang H, He L, Zhao J-j, Coppola D, Dalton WS, Cheng JQ: **MicroRNA-155 is regulated by the transforming growth factor beta/Smad pathway and contributes to epithelial cell plasticity by targeting RhoA.** *Mol Cell Biol* 2008, **28**:6773-6784.
69. Ma L, Teruya-Feldstein J, Weinberg RA: **Tumour invasion and metastasis initiated by microRNA-10b in breast cancer.** *Nature* 2007, **449**:682-688.
70. Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, Becker KG, Gorospe M, Hide W, Lieberman J: **miR-24 Inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3'UTR microRNA recognition elements.** *Mol Cell* 2009, **35**:610-625.
71. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO, Tavaré S, Caldas C, Miska EA: **MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.** *Genome Biol* 2007, **8**:R214.
72. Kota J, Chivukula RR, O'Donnell KA, Wentzel EA, Montgomery CL, Hwang HW, Chang TC, Vivekanandan P, Torbenson M, Clark KR, Mendell JR, Mendell JT: **Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model.** *Cell* 2009, **137**:1005-1017.
73. Mayr C, Hemann MT, Bartel DP: **Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation.** *Science* 2007, **315**:1576-1579.
74. Wu L, Fan J, Belasco JG: **MicroRNAs direct rapid deadenylation of mRNA.** *Proc Natl Acad Sci USA* 2006, **103**:4034-4039.
75. Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, Labourier E, Reinert KL, Brown D, Slack FJ: **RAS is regulated by the let-7 microRNA family.** *Cell* 2005, **120**:635-647.
76. Christoffersen NR, Shalgi R, Frankel LB, Leucci E, Lees M, Klausen M, Pilpel Y, Nielsen FC, Oren M, Lund AH: **p53-independent upregulation of miR-34a during oncogene-induced senescence represses MYC.** *Cell Death Differ* 2010, **17**:236-245.
77. Welch C, Chen Y, Stallings RL: **MicroRNA-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells.** *Oncogene* 2007, **26**:5017-5022.
78. Sun F, Fu H, Liu Q, Tie Y, Zhu J, Xing R, Sun Z, Zheng X: **Downregulation of CCND1 and CDK6 by miR-34a induces cell cycle arrest.** *FEBS Lett* 2008, **582**:1564-1568.
79. Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ, Gimotty PA, Katsaros D, Coukos G, Zhang L, Puré E, Agami R: **The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis.** *Nat Cell Biol* 2008, **10**:202-210.

doi:10.1186/bcr2484
Cite this article as: O'Day E, Lal A: **MicroRNAs and their target gene networks in breast cancer.** *Breast Cancer Research* 2010, **12**:201.

Molecular Cell
Article

miR-24 Inhibits Cell Proliferation by Targeting E2F2, MYC, and Other Cell-Cycle Genes via Binding to "Seedless" 3′UTR MicroRNA Recognition Elements

Ashish Lal,[1,*] Francisco Navarro,[1] Christopher A. Maher,[3] Laura E. Maliszewski,[1] Nan Yan,[1] Elizabeth O'Day,[1] Dipanjan Chowdhury,[1,2] Derek M. Dykxhoorn,[1,7] Perry Tsai,[1] Oliver Hofmann,[5] Kevin G. Becker,[4] Myriam Gorospe,[4] Winston Hide,[5,6] and Judy Lieberman[1,*]

[1]Immune Disease Institute, Program in Cellular and Molecular Medicine, Children's Hospital Boston and Department of Pediatrics
[2]Dana Farber Cancer Institute and Department of Radiation Oncology
Harvard Medical School, Boston, MA 02115, USA
[3]Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI 48109, USA
[4]Research Resources Branch and Laboratory of Cellular and Molecular Biology, National Institute on Aging-IRP, National Institutes of Health, Baltimore, MD 21224, USA
[5]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
[6]South African National Bioinformatics Institute, University of the Western Cape, Bellville 7535, South Africa
[7]The John T. Macdonald Foundation Department of Human Genetics and the Department of Microbiology and Immunology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA
*Correspondence: alal@idi.harvard.edu (A.L.), lieberman@idi.harvard.edu (J.L.)
DOI 10.1016/j.molcel.2009.08.020

SUMMARY

miR-24, upregulated during terminal differentiation of multiple lineages, inhibits cell-cycle progression. Antagonizing miR-24 restores postmitotic cell proliferation and enhances fibroblast proliferation, whereas overexpressing miR-24 increases the G1 compartment. The 248 mRNAs downregulated upon miR-24 overexpression are highly enriched for DNA repair and cell-cycle regulatory genes that form a direct interaction network with prominent nodes at genes that enhance (*MYC*, *E2F2*, *CCNB1*, and *CDC2*) or inhibit (*p27Kip1* and *VHL*) cell-cycle progression. miR-24 directly regulates *MYC* and *E2F2* and some genes that they transactivate. Enhanced proliferation from antagonizing miR-24 is abrogated by knocking down *E2F2*, but not *MYC*, and cell proliferation, inhibited by miR-24 overexpression, is rescued by miR-24-insensitive *E2F2*. Therefore, *E2F2* is a critical miR-24 target. The *E2F2* 3′UTR lacks a predicted miR-24 recognition element. In fact, miR-24 regulates expression of *E2F2*, *MYC*, *AURKB*, *CCNA2*, *CDC2*, *CDK4*, and *FEN1* by recognizing seedless but highly complementary sequences.

INTRODUCTION

MicroRNAs (miRNAs) regulate key steps of cell differentiation and development by suppressing gene expression in a sequence-specific manner (Bartel, 2009). In mammals, the active strand miRNA sequence (typically ~22 base pairs) is partially complementary to binding sites in the 3′UTR of genes, often with full complementarity to 7 or 8 nucleotides in the "seed region" (residues 2–9) of the miRNA. Gene suppression in mammals is thought to occur primarily by inhibiting translation (Olsen and Ambros, 1999). However, miRNAs in mammals also cause mRNA decay (Chang et al., 2007; Lim et al., 2005; Johnson et al., 2007); recent reports (Baek et al., 2008; Selbach et al., 2008) suggest that reduced protein is frequently associated with decreased mRNA.

miR-24 is consistently upregulated during terminal differentiation of hematopoietic cell lines into a variety of lineages (Lal et al., 2009). miR-24 is also upregulated during thymic development to naive CD8 T cells (Neilson et al., 2007) and during muscle and neuronal cell differentiation (Sun et al., 2008; Fukuda et al., 2005). miR-24 is encoded with miR-23 and miR-27 in two duplicated gene clusters. One cluster (miR-23b, miR-27b, and miR-24-1) is within a chromosome 9 EST, and the other (miR-23a, miR-27a, and miR-24-2) is in a chromosome 19 intergenic region. Both miR-24 genes are processed to the same active strand. Disruption or changes in expression of both sites have been linked to CLL prognosis (Calin et al., 2005). Because miR-24 is upregulated in diverse cell types during terminal differentiation, we sought to identify its function and the target genes that it regulates.

Common approaches to identify miRNA target genes are (1) bioinformatic algorithms that predict potential target genes that contain conserved 3′UTR sequences complementary to a seed region at the 5′ end of the miRNA active strand (Doench and Sharp, 2004; Lewis et al., 2005), (2) analysis of mRNAs that are downregulated when a miRNA is overexpressed (Chang et al., 2007; Johnson et al., 2007; Lim et al., 2005), and (3) identifying mRNAs enriched in coimmunoprecipitates with tagged Argonaute or GW182 proteins in cells overexpressing the miRNA (Easow et al., 2007; Zhang et al., 2007). The bioinformatic approach is hampered by the fact that the existing algorithms have a high margin of error (most predicted genes are not real targets, and some key targets, *such as RAS* for let-7, are not predicted [Johnson et al., 2005]). The utility of the biochemical approach involving Argonaute proteins for genome-wide target

identification of miRNAs is still unclear because Argonaute overexpression globally increases miRNA levels, perhaps obscuring the effect of an individual overexpressed miRNA (Diederichs and Haber, 2007). Because miRNA-mediated mRNA degradation and protein downregulation often occur together (Baek et al., 2008), identifying the mRNAs that decrease when a miRNA is overexpressed might identify many of its targets. Although some bona fide miR-24 targets that are primarily regulated by translation will be missed by this approach and other downregulated genes may not be directly regulated, this strategy has been successfully used to identify targets of some mammalian miRNAs, including miR-124 and miR-1 (Lim et al., 2005), miR-34a (Chang et al., 2007), and let-7 (Johnson et al., 2007). Therefore, we applied this approach to identify the genes regulated by miR-24 in HepG2 cells that express low levels of miR-24 and combined it with bioinformatics to uncover miR-24-regulated pathways. We find that miR-24 regulates a network of genes that control cell-cycle progression and DNA repair (Lal et al., 2009). Overexpressing miR-24 increases the G1 population and reduces DNA replication, whereas antagonizing miR-24 increases cell proliferation, which can be rescued by knocking down *E2F2*, suggesting that *E2F2* is a key miR-24 target gene. *MYC* and other genes important in cell-cycle regulation that are transcriptionally regulated by MYC and E2Fs (*AURKB*, *BRCA1*, *CCNA2*, *CDC2*, *CDK4*, and *FEN1*) are also direct miR-24 targets by luciferase assay. Of note, *E2F2* and most of these genes lack 3′UTR miR-24 seed match sequences. However, miR-24 regulates these genes by base pairing to "seedless" 3′UTR MREs with extensive base pairing elsewhere in the sequence.

RESULTS

miR-24 Is Upregulated during Hematopoietic Differentiation

To understand the role of miRNAs during terminal differentiation, we analyzed miRNA expression by microarray in two human leukemia cell lines: K562 cells differentiated to megakaryocytes using 12-O-tetradecanoylphorbol-13-acetate (TPA) or to erythrocytes with hemin and HL60 cells differentiated to macrophages using TPA or to monocytes using vitamin D3. miR-24 was one of only six miRNAs that was consistently upregulated in all four systems of terminal differentiation (Lal et al., 2009). The other uniformly upregulated miRNAs were three other members of the miR-24 clusters (miR-23a, miR-23b, and miR-27a), miR-22, and miR-125a. miR-24 was the most upregulated of these miRNAs. We therefore focused on miR-24, which we hypothesized might regulate terminal differentiation in multiple cell lineages. qRT-PCR confirmed the induction of miR-24 during differentiation of these hematopoietic cells (Figure 1A) with the highest upregulation in K562 cells treated with TPA. The mature miR-24 transcript increased 2- to 8-fold during differentiation into megakaryocytes, erythrocytes, macrophages, monocytes, and granulocytes. Expression of the chromosome 19 miR-24 cluster primary transcript encoding miR-23a, miR-27a, and miR-24 increased in both cell lines within 6 hr of TPA treatment, peaked at ~12 hr, and remained elevated for at least 2 days (Figures S1A and S1B available online), suggesting that the observed increase in mature miR-24 was due to increased transcription. Upregulation of the Dicer-cleaved mature miRNA was slightly delayed, becoming significant at 12–16 hr (Figures S1C and S1D). Mature miR-24 levels remained elevated for as long as was measured (4 days).

miR-24 Inhibits Cellular Proliferation by Increasing the G1 Compartment

Because cessation of cell proliferation is a hallmark of terminal differentiation, we first examined whether proliferation is altered by either inhibiting or enhancing miR-24 function by transfecting cells with miR-24 2′-OMe antisense oligonucleotide (ASO) or miRNA mimics, respectively. When K562 cells were transfected with miR-24 ASO, miR-24 was dramatically and specifically reduced by qRT-PCR 36 hr later (Figure 1B). DNA replication, measured by thymidine incorporation, doubled in cells transfected with miR-24 ASO compared to cells transfected with control ASO (Figure 1C). When K562 cells were differentiated with TPA for 4 hr, thymidine incorporation declined by 60%. However, in cells transfected with miR-24 ASO and treated with TPA, thymidine uptake was indistinguishable from that of the control ASO-transfected, but TPA-untreated, cells (Figure 1C). Therefore, miR-24 ASO fully restored proliferation to differentiating K562 cells. To examine whether miR-24 also inhibits cell proliferation in nontransformed cells, we next antagonized miR-24 in early passage WI-38 and IMR-90 normal diploid fibroblasts. Antagonizing miR-24 in WI-38 and IMR-90 cells dramatically reduced miR-24 (Figure 1D) and increased thymidine uptake > 2-fold 48 hr after transfection (Figure 1E). Conversely, overexpressing miR-24 in HepG2 cells synchronized with nocodazole, which typically leads to mitotic arrest and only ~8% of cells in G1, increased G1 cells 3-fold (22%; miR-24 versus cel-miR-67, $p < 0.001$) (Figure 1F).

We next analyzed how miR-24 expression changes during normal cell-cycle progression using K562 cells released at various times from nocodazole treatment, which synchronized them in G2/M (Bar-Joseph et al., 2008; O'Donnell et al., 2005) (Figures 1G and 1H). Before release, 90% of cells were in G2/M; 8 hr later, 65% were in G1; and 12 hr after removing nocodazole, 45% were in S phase. miR-24 was low in G2/M, increased > 3-fold by 8 hr when most cells were in G1, and then declined by 12 hr as cells progressed into S phase. These results suggest that miR-24 is most highly expressed in G1. Taken together with our finding that cells transduced with miR-24 mimics accumulate in G1, these results suggest that miR-24 regulates cell-cycle progression mostly by blocking or delaying the G1/S transition.

Most Genes Downregulated by miR-24 Contain a miR-24 Seed in Their 3′UTR

We next sought to identify miR-24-regulated targets and cellular pathways by comparing mRNA microarrays of cells transfected with miR-24 or control miRNA (cel-miR-67) mimic. Transfecting HepG2 cells, which have low endogenous miR-24 levels (Figure 2A), with a miR-24 mimic increased miR-24 expression ~80-fold compared with control cells (Figure 2B). Total RNA, isolated from duplicate miRNA-transfected samples 48 hr later, was amplified, labeled, and hybridized to Illumina mRNA microarrays. 248 mRNAs were downregulated at least 2-fold by miR-24 overexpression (Z ratio > 1.5) (Table S1). We validated the microarray

A

Cell	Differentiating agent	Differentiated to	Fold increase in miR-24
K562	None	-	1.0 ± 0.2
K562	TPA	Megakaryocytes	8.0 ± 1.2
K562	Hemin	Erythrocytes	3.7 ± 0.4
HL60	None	-	1.0 ± 0.2
HL60	TPA	Macrophages	4.8 ± 0.5
HL60	VitD3	Monocytes	2.3 ± 0.3
HL60	DMSO	Granulocytes	2.9 ± 0.4

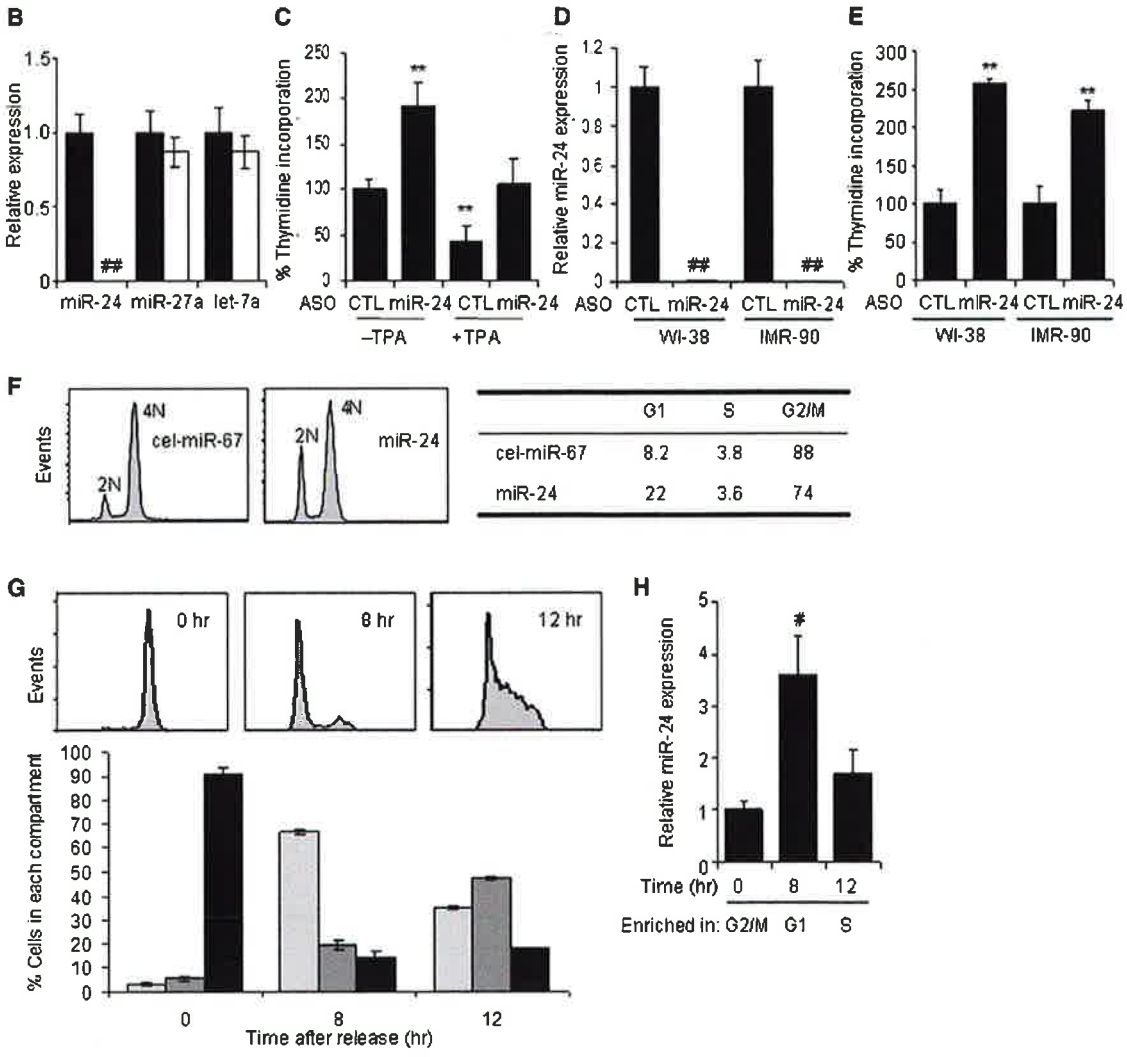


Figure 1. miR-24 Is Upregulated during Hematopoietic Cell Differentiation and Inhibits Cell Proliferation

(A) miR-24 expression, measured using qRT-PCR relative to untreated cells, increases in K562 cells differentiated to megakaryocytes or erythrocytes and HL60 cells differentiated to macrophages, monocytes, and granulocytes. Differentiation in all experiments was verified by cell surface phenotype. Mature miR-24 levels were determined by qRT-PCR and normalized to *U6*.

(B) miR-24 knockdown in K562 cells specifically decreases miR-24, assayed by qRT-PCR in cells transfected with miR-24 ASO (white) relative to control ASO (black). Expression relative to *U6* snRNA is normalized to control cells.

(C) miR-24 knockdown with ASO increases K562 cell proliferation measured by thymidine uptake, both in the presence and absence of TPA. The decline in proliferation with TPA is completely restored by antagonizing miR-24.

data by performing qRT-PCR amplification for nine randomly chosen downregulated genes that spanned the range of significantly downregulated genes (Z ratios, 1.6–6.1), of which six (*CNDP2*, *TOP1*, *PER2*, *MBD6*, *H2AFX*, and *STX16*) were predicted miR-24 targets by TargetScan 4.2 and three (*UBD*, *BCL2L12*, and *ZNF317*) were not (Figure 2C). *H2AFX* was previously shown to be directly regulated by miR-24 (Lal et al., 2009). All nine genes were significantly downregulated, and the extent of downregulation correlated well with their reduced expression in the microarray, validating the quality of the microarray data.

To determine what fraction of the downregulated genes are likely direct targets of miR-24, we used two approaches. First, we compared our experimental list of downregulated transcripts with TargetScan predictions. One hundred downregulated genes were also predicted by TargetScan 4.2 (Lewis et al., 2003) (Figure 2D and Table S1). Among these 100 targets, however, only 20 have predicted miR-24 3′UTR miRNA recognition sites (MRE) conserved in human, mouse, rat, and dog. Second, we examined the frequency of a 3′UTR sequence perfectly complementary to the miR-24 seed (hexamer [positions 2–7], heptamer [positions 2–8], and octamer [positions 2–9]). The downregulated genes were highly enriched for miR-24 seed matches (Figure 2E and Table S1). Just more than half of the 219 miR-24-downregulated transcripts that have an annotated 3′UTR contain a 3′UTR complementary hexamer sequence (53%, $p = 2 \times 10^{-16}$ relative to the background frequency of the seed in the known transcriptome); 32% have a heptamer match ($p = 7 \times 10^{-15}$); and 8% have an octamer seed match ($p = 0.0002$). This significant enrichment of predicted miR-24 target genes and the high frequency of genes containing perfect seed matches suggest that a substantial proportion of the downregulated genes may be direct miR-24 targets.

Cell Cycle and DNA Repair Genes Are Regulated by miR-24

We next performed a gene ontology (GO) analysis (GeneGo, Inc.) to identify cellular pathways enriched in the set of genes downregulated after ectopic miR-24 expression. This analysis did not yield a tightly focused set of functions for miR-24 (Table S2). Therefore, we next looked at whether the 100 downregulated genes, which were also predicted miR-24 targets, are enriched for specific biological processes. A functional enrichment and network analysis revealed statistically significant enrichment for 49 processes, many of which are overlapping (Figure 3A and Table S3). The top three most enriched GO processes involve DNA repair (DNA damage checkpoint, double-strand break repair by homologous recombination, and recombinational repair; each enriched with a significance of $p = 0.0001$). We previously found that miR-24 interferes with the DNA damage response in terminally differentiated hematopoietic cells, predominantly by reducing expression of the histone variant H2AFX, which recruits and retains DNA repair factors at double-strand breaks (Lal et al., 2009). In addition, multiple GO processes involved in cell-cycle regulation were also highly enriched (regulation of cell cycle, $p = 0.0002$; DNA integrity checkpoint, $p = 0.0003$; cell-cycle arrest, $p = 0.0007$; cell cycle, $p = 0.001$; DNA recombination, $p = 0.001$). This was not surprising based on the effect of miR-24 on cell-cycle progression. When networks were developed to identify known directly interacting proteins from these overrepresented biological processes, there was one cluster of six genes centered around *MYC* (*c-myc*) and three other small clusters involving two or three genes (Figure S2A); the other 87 genes in the data set lack any previously annotated direct interactions. Because the TargetScan algorithm might miss some important miRNA-regulated genes, we also constructed a direct interaction network from the 248 downregulated mRNAs. The direct interaction network constructed from all significantly downregulated mRNAs was a highly interactive set of 68 interacting genes, many of which are important in cell-cycle regulation. The major connected network of miR-24-downregulated genes is shown in Figure 3B; there were also some smaller networks (Figure S2B). Key nodes of the major network are *MYC* (22 interactions), *E2F2* (six interactions), *VHL* (six interactions), *CDC2* (six interactions), *CCNB1* (five interactions), and *CDKN1B* (five interactions). The MYC and E2F2 transcription factors play a central role in regulating G1/S transition and progression through S. They inhibit cell differentiation and apoptosis and promote cellular transformation (Bracken et al., 2004; Lebofsky and Walter, 2007). MYC regulates the transcription of other genes in the network, including *E2F2*, *CDKN1B*, *CCNB1*, *CDC2*, *CDCA7*, and *RRM2*. E2F2 also regulates the transcription of other genes in the network, including *CDC2*, *MYC*, *RRM2*, and the mini-chromosome maintenance proteins MCM4 and MCM10 that are essential for initiating DNA replication. These analyses support our experimental findings that miR-24 regulates cell-cycle progression and DNA repair.

miR-24 Regulates *MYC* by Binding to Its 3′UTR

To determine whether miR-24 directly regulates the downregulated genes, we began with *MYC* because it is a key node of the interaction network, plays an important role in cell-cycle progression, and is a predicted miR-24 target. We transfected HepG2 and K562 cells with miR-24 mimics and, 48 hr later, measured *MYC* mRNA by qRT-PCR. miR-24 overexpression decreased *MYC* mRNA by ~2- to ~4-fold relative to *GAPDH* in

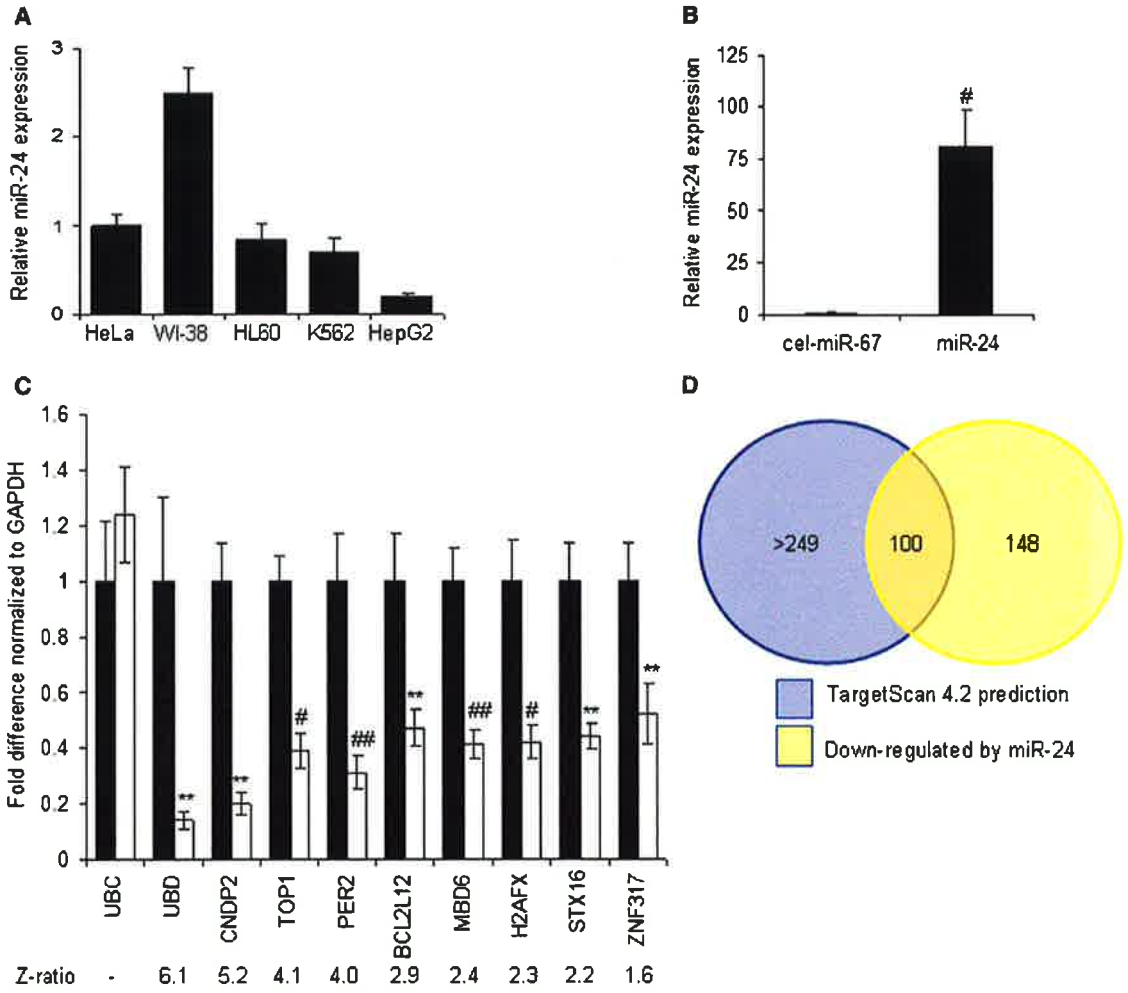
(D and E) Knocking down miR-24 in WI-38 or IMR-90 cells also significantly increases proliferation as measured by thymidine incorporation 48 hr posttransfection. miR-24 knockdown by ASO measured by qRT-PCR is normalized to *U6* snRNA.

Error bars in (B–E) represent SD from three experiments. $^{**}p < 0.01$, $^{\#\#}p < 0.001$.

(F) miR-24 overexpression increases the G1 compartment in HepG2 cells. HepG2 cells transfected with miR-24 or control mimic for 48 hr were stained with propidium iodide and analyzed by flow cytometry. Representative analysis of three independent experiments is shown.

(G) K562 cells, synchronized in G2/M by nocodazole and then released, were analyzed by flow cytometry. A representative experiment is shown in the top panel, and the mean (± SD) percentage of cells in each phase of the cell cycle (from three independent experiments) is shown below (light gray, G1; dark gray, S; black, G2/M).

(H) qRT-PCR analysis of miR-24 (normalized to *U6* snRNA) from the partially synchronized K562 cells in (G) shows that miR-24 is most highly expressed in G1 and declines as cells progress to S and G2/M phase. $^{\#}p < 0.005$.



E

Target motif	Motif length	miR-24 match nucleotides	% miR-24 down-regulated genes with motif	p-value
UGAGCC	6	2-7	53	2.2×10^{-16}
CUGAGCC	7	2-8	32	6.8×10^{-15}
ACUGAGCC	8	2-9	8	1.7×10^{-4}

Figure 2. mRNA Downregulation after miR-24 Overexpression

(A) HepG2 cells express low levels of miR-24, assayed by qRT-PCR analysis normalized to *U6*, compared to HeLa, WI-38, HL60, and K562 cells.

(B) Effective increase in miR-24 in HepG2 cells 48 hr after transfection with miR-24 mimic compared to cel-miR-67-transfected control cells. $^{\#}p < 0.005$.

(C) Genes identified by microarray as downregulated by miR-24 overexpression were confirmed to be downregulated by qRT-PCR normalized to *GAPDH*. *UBC* is a housekeeping gene. Cells were transfected with cel-miR-67 (black) or miR-24 mimic (white). $^{**}p < 0.01$, $^{\#}p < 0.005$, and $^{\#\#}p < 0.001$. The downregulated genes are graphed in order of their downregulation on the microarray; the Z ratio of the microarray analysis is shown below. Error bars represent SD from three independent experiments (A, B, and C).

(D) Venn diagram of genes downregulated by miR-24 in HepG2 cells and genes predicted to be regulated by miR-24 using TargetScan 4.2. Of the 100 predicted genes, whose mRNA is also significantly downregulated, only 20 have conserved predicted miR-24 recognition sites. TargetScan 4.2 predicts 349 conserved miR-24 targets and many more that are not conserved.

(E) Sites complementary to the miR-24 seed are enriched in the 3′UTR of downregulated transcripts. The table shows the frequency of perfect hexamer (positions 2–7), heptamer (positions 2–8), and octamer (positions 2–9) miR-24 3′UTR seeds in the downregulated genes.

HepG2 and K562 cells (Figures 4A and 4B). *UBC* mRNA, a control transcript, did not change significantly. MYC protein also decreased substantially (85%) (Figure 4C). To investigate whether reduced MYC expression was directly mediated by miR-24, we measured changes after miR-24 cotransfection in luciferase activity from a *MYC* 3′UTR reporter. miR-24 reduced luciferase activity 2.2-fold (Figure 4D). We next sought to identify miR-24 MREs in the *MYC* 3′UTR. TargetScan 4.2 predicts a single MRE containing a poorly conserved 7-mer exact seed match at positions 462–468 (although the recent TargetScan 5.0 algorithm does not list *MYC* as a miR-24 target), whereas rna22, an algorithm that does not require a seed match (Miranda et al., 2006), identifies six potential miR-24 MREs in the 488 nt *MYC* 3′UTR, including the TargetScan 4.2-predicted MRE (MRE6) (Figure 4E). miR-24 overexpression specifically and significantly reduced luciferase activity by 1.9- and 3.9-fold for MRE3 and MRE6, respectively, but the other MREs had no significant effect on luciferase activity (Figure 4F). MRE3 has no seed but has extensive complementarity with the 3′ end of miR-24. Point mutations of MRE3 and MRE6 that disrupted miR-24 binding restored luciferase activity (Figures 4G and 4H). These findings suggest that miR-24 binds to two partially complementary sites (MRE3 and MRE6) in the *MYC* 3′UTR.

miR-24 Downmodulates *E2F2*

We next examined the effect of miR-24 on *E2F2* because *E2F2* is downregulated by miR-24 overexpression by microarray, is a key node in the gene interaction network (Figure 3B), and plays a crucial role in regulating progression through G1, where miR-24-overexpressing cells pile up (Polager and Ginsberg, 2008). qRT-PCR analysis confirmed that *E2F2* mRNA was significantly downregulated by overexpressing miR-24 (Figure 5A). In addition, the related E2F family members, *E2F1* and *E2F3*, were also significantly decreased, although these two genes were not identified by the less sensitive microarray analysis. *E2F1* and *E2F3* downregulation may be secondary to *E2F2* downregulation because the E2F family of transcription factors regulates each other (Bracken et al., 2004; Vernell et al., 2003) or may be mediated by MYC, given that MYC and E2F1 have been shown to transactivate each other (Fernandez et al., 2003). As expected, E2F2 protein (Figure 5B) was also substantially reduced (9-fold).

miR-24 Downregulates Multiple *E2F*- and *MYC*-Regulated Genes

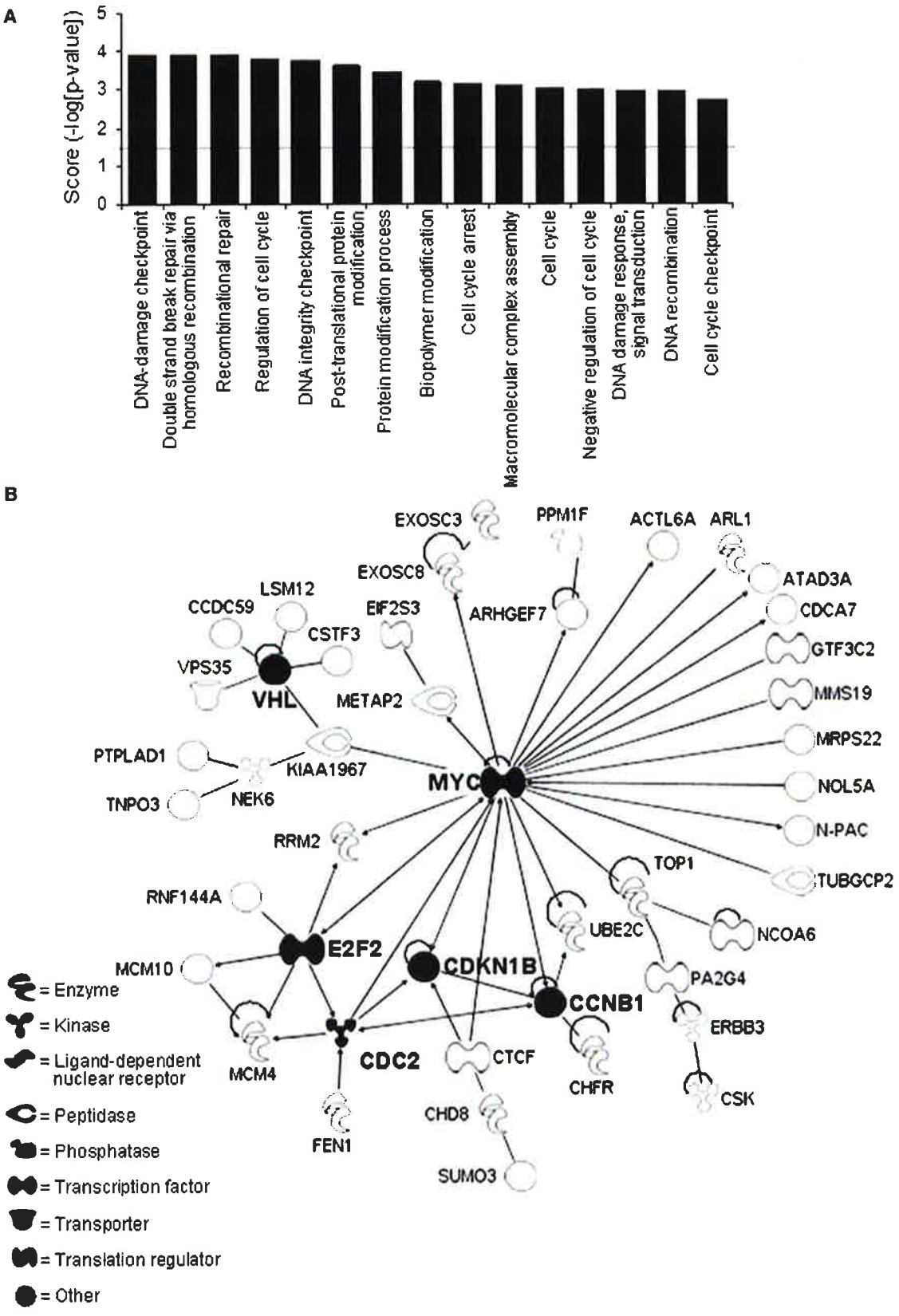
The E2F transcription factors activate the transcription of many genes essential for DNA replication, cell-cycle progression, and DNA repair. If miR-24 overexpression downregulates *E2F2*, *E2F2* target gene mRNAs would also be expected to decline after ectopic miR-24 expression. The effect of ectopic miR-24 expression in HepG2 cells on transcripts of 10 E2F targets that are important for cell-cycle progression and DNA repair (*AURKB*, *BRCA1*, *CCNA2*, *CDC2*, *CHEK1*, *FEN1*, *PCNA*, *RRM2*, *MCM4*, and *MCM10*) was analyzed 48 hr later. Eight of the ten transcripts, with the exception of *BRCA1* and *PCNA*, were significantly reduced (>40%) (Figure 5A). A subset of these genes (*CDC2*, *MCM4*, *MCM10*, *RRM2*, and *FEN1*) was also significantly downregulated by miR-24 overexpression by microarray (Table S1). mRNA microarray may not be sensitive enough to identify some genes whose expression is suppressed, either directly or indirectly, by a miRNA. Protein levels of E2F2 and all seven miR-24 target genes examined (AURKB, BRCA1, CCNA2, CDC2, CHEK1, FEN1, and PCNA), quantified by densitometry of immunoblots, decreased by at least 2-fold (Figure 5B). A possible explanation for the lack of correlation between the mRNA and protein levels of BRCA1 and PCNA is that mRNA levels were measured 48 hr after transfection of one cell type (HepG2), whereas protein levels were assayed 72 hr posttransfection in K562 cells. In fact, *BRCA1*, but not *PCNA*, mRNA was reduced by ~2-fold when K562 cells were transfected with miR-24 for 3 days (Figure S5).

Because miR-24 overexpression reduced MYC protein by 85% (Figure 4C), MYC target genes should also be downregulated. Consistent with this hypothesis, 11 known MYC-regulated genes (*ATAD3A*, *ACTL6A*, *ARHGEF7*, *CCNB1*, *CDCA7*, *EXOSC8*, *E2F2*, *METAP2*, *N-PAC*, *RRM2*, and *UBE2C*) were significantly downregulated in miR-24-overexpressing HepG2 cells by mRNA microarray (Table S1). Among MYC-regulated genes, CDK4 is an important mediator of MYC's effects on cellular proliferation (Hermeking et al., 2000). Although *CDK4* mRNA was not significantly altered by microarray, *CDK4* mRNA declined ~4-fold after ectopic expression of miR-24 in HepG2 cells by more sensitive qRT-PCR assay (Figure 5A), and CDK4 protein became undetectable (Figure 5B). Therefore, miR-24 overexpression decreases the levels of many genes that are important in cell-cycle progression.

In these experiments, we overexpressed miR-24 ~80-fold above the level in undifferentiated K562 cells, whereas the physiological increase after TPA treatment of K562 cells is only 8-fold. To determine whether these genes are regulated by a physiological increase in miR-24, these experiments were repeated by transfecting K562 cells with varying miR-24 mimic concentrations (2–50 nM). Transfection of 2 nM miR-24 did not significantly alter miR-24, whereas 10 and 50 nM miR-24 increased miR-24 levels by 4- and 28-fold, respectively (Figure 5C). *E2F2*, *MYC*, and three of four other E2F2-regulated mRNAs (*AURKB*, *CCNA2*, and *H2AX*, but not *PCNA*) (Figure S5) and protein levels of all seven genes tested (*E2F2*, *AURKB*, *CHEK1*, *CCNA2*, *CDK4*, *MYC*, and *PCNA*) were all significantly reduced by a 4-fold increase in miR-24 (Figure 5D). Therefore, the genes that were identified by mRNA microarray as downregulated after ectopic miR-24 expression are likely physiologically relevant direct and/or indirect miR-24 targets.

E2F2 and Some E2F Target Genes Are Directly Regulated by Seedless 3′UTR MREs

None of the three *E2F* paralogs (*E2F1*, *E2F2*, and *E2F3*) are a predicted target of miR-24, and their 3′UTRs do not contain a miR-24 seed match sequence. We nonetheless tested whether the *E2F* 3′UTRs might be directly regulated by miR-24 by luciferase assay. The *E2F2*, but not *E2F1* or *E2F3*, 3′UTR significantly repressed luciferase activity in a miR-24-dependent manner (Figure 6A), suggesting that *E2F2* is a direct miR-24 target. rna22 identified five candidate *E2F2* 3′UTR miR-24 MREs (Figures 6B and S3). miR-24 significantly suppressed luciferase activity of a reporter gene containing the *E2F2* MRE1. *E2F2* MRE1 does not have a seed match in the 3′UTR, even if G:U


A
Score (-log[p-value])
0
1
2
3
4
5
DNA-damage checkpoint
Double strand break repair via homologous recombination
Recombinational repair
Regulation of cell cycle
DNA integrity checkpoint
Post-translational protein modification
Protein modification process
Biopolymer modification
Cell cycle arrest
Macromolecular complex assembly
Cell cycle
Negative regulation of cell cycle
DNA damage response, signal transduction
DNA recombination
Cell cycle checkpoint
B
EXOSC3
PPM1F
ACTL6A
ARL1
EXOSC8
ATAD3A
LSM12
CCDC59
EIF2S3
ARHGEF7
CDCA7
CSTF3
GTF3C2
VPS35
VHL
METAP2
MMS19
MRPS22
PTPLAD1
KIAA1967
MYC
NOL5A
TNPO3
NEK6
N-PAC
RRM2
TOP1
TUBGCP2
RNF144A
UBE2C
NCOA6
E2F2
CDKN1B
MCM10
PA2G4
CCNB1
ERBB3
MCM4
CDC2
CTCF
CHFR
CSK
CHD8
FEN1
SUMO3
= Enzyme
= Kinase
= Ligand-dependent nuclear receptor
= Peptidase
= Phosphatase
= Transcription factor
= Transporter
= Translation regulator
= Other

wobbles are allowed, but has extensive complementarity to miR-24 elsewhere. Point mutations that disrupt base pairing between miR-24 and *E2F2* MRE1 rescued luciferase expression, verifying that miR-24 specifically recognizes the *E2F2* MRE1.

To verify further that *MYC* and *E2F2* are direct targets of miR-24, we also looked at changes in expression of luciferase reporter genes 24 hr after transduction of HepG2 cells with miR-24 mimic. At this early time, thymidine uptake of HepG2 cells does not significantly change (Figure S4A), but ectopic miR-24 still suppresses luciferase reporters encoding *MYC* MRE3 or MRE6 or *E2F2* MRE1 (Figure S4B). The identification of seedless *E2F2* and *MYC* MREs confirms previous studies showing that MREs lacking a seed with good downstream complementarity can contribute to miRNA gene regulation (Didiano and Hobert, 2006, 2008; Vella et al., 2004).

Because seedless MREs contributed to the regulation of *E2F2* and *MYC* by miR-24, we next investigated whether some of the *E2F2*- and *MYC*-regulated genes, whose transcripts declined in response to miR-24, might also be direct miR-24 targets even though they might lack a predicted MRE. We selected eight genes (*AURKB*, *BRCA1*, *CCNA2*, *CHEK1*, *CDC2*, *CDK4*, *FEN1*, and *PCNA*) that play important roles in cell-cycle progression and cloned their entire 3′UTRs into the luciferase reporter. The 3′UTR of six out of eight genes (*AURKB*, *BRCA1*, *CCNA2*, *CDC2*, *CDK4*, and *FEN1*, but not *CHEK1* or *PCNA*) was significantly repressed by miR-24, suggesting that these genes may be direct targets (Figure 6C). To confirm that these genes are direct miR-24 targets, we next sought to identify the miR-24 MREs that regulate their expression. Among the six genes whose 3′UTR was specifically repressed by miR-24, *BRCA1* is the only gene that is predicted by TargetScan. The *BRCA1* 3′UTR contains a nonconserved perfect 7-mer seed match sequence that functions as a miR-24 MRE by luciferase assay (Figures 6D and 6E). To identify potential miR-24 MREs in these *E2F2*- and *MYC*-regulated genes, we used the rna22 or PITA algorithms, which allow G:U wobbles or seed mismatches. These algorithms identified one candidate MRE for *AURKB*; five for *BRCA1* (which included the TargetScan *BRCA1* site); and three sites each for *CCNA2*, *CDC2*, *CDK4*, and *FEN1* (Figure S6). miR-24 significantly repressed luciferase activity of one MRE for five out of six of these reporter genes (*AURKB* MRE1, *BRCA1* MRE5, *CDC2* MRE1, *CDK4* MRE1, and *FEN1* MRE1) (Figures 6D, 6E, and S7). Although *CCNA2* MRE1 appeared to be inactive, a longer fragment (181 nucleotides) from the *CCNA2* 3′UTR (that included only *CCNA2* MRE1) significantly repressed luciferase expression in a miR-24-dependent manner when cloned into the luciferase vector 3′UTR (Figure 6F). Point mutations that disrupt base pairing between miR-24 and the five minimal MREs and the *CCNA2* MRE within the extended sequence rescued luciferase expression, verifying that these MREs are regulated by miR-24 (Figures 6D–6F). Therefore, we have identified and verified by mutation seven seedless miR-24 MREs in genes important in cell-cycle progression.

E2F2 Downregulation Is Key to miR-24's Inhibition of Cell Proliferation

Because both MYC and E2F2 are important cell-cycle progression regulators, we next examined their contributions to the increased cellular proliferation from antagonizing miR-24 by knocking down *MYC* and/or *E2F2* in K562 cells cotransfected with miR-24 ASO (Figures 7A and S8). Introducing miR-24 ASO into K562 cells doubled thymidine incorporation (as in Figure 1C). *E2F2* knockdown completely abrogated the proliferative effect of miR-24 ASO, but *MYC* knockdown had no significant effect. Moreover, *E2F2* downregulation by miR-24 is physiologically relevant. When K562 cells were terminally differentiated to megakaryocytes with TPA, the decrease in *E2F2* mRNA and protein was completely blocked by inhibiting miR-24 (Figures 7B and 7C). Conversely, ectopic expression of miR-24-insensitive *E2F2* lacking the 3′UTR restored proliferation to miR-24-treated K562 cells (Figures 7D and 7E). Therefore, the miR-24 antiproliferative effect is largely mediated by its downregulation of *E2F2*.

Antagonizing miR-24 elevated MYC protein levels in untreated K562 cells, and the downregulation of *MYC* mRNA in TPA-treated K562 cells could be partially rescued by antagonizing miR-24 (Figures S9A and S9B). However, antagonizing miR-24 did not restore MYC protein to differentiating cells, suggesting that, although miR-24 suppresses *MYC* expression, downregulation of MYC protein during postmitotic differentiation is also controlled by miR-24-independent changes in protein stability. This may help to explain why *MYC* siRNAs had no significant effect on proliferation of cells transduced with miR-24 ASO (Figure 7A).

DISCUSSION

miR-24 and its clustered miRNAs are among only a handful of miRNAs consistently upregulated during hematopoietic cell terminal differentiation. Here, we show that miR-24 suppresses expression of several key genes that regulate cell-cycle progression. Overexpressing miR-24 increases the percentage of cells in the G1 phase, whereas antagonizing it causes differentiating cells to keep proliferating. The antiproliferative effect of miR-24 is not restricted to tumor cells (HepG2 and K562 cells) but also occurs in human diploid fibroblasts.

miRNAs can regulate expression of hundreds of genes. Genome-wide analysis of miRNA target genes has been assessed following miRNA overexpression or knockdown for only a handful of miRNAs (Chang et al., 2007; Johnson et al., 2007; Lim et al., 2005). Using this approach for miR-24 enabled us to identify

Figure 3. Bioinformatic Analysis of miR-24-Downregulated Genes Suggests that miR-24 Regulates Cell-Cycle Progression and DNA Repair

(A) Top 15 overrepresented cellular processes for the 100 overlapping genes in Figure 2D. Overrepresented processes are sorted by score (−log [p value]). A highly positive score suggests that the subnetwork is saturated with genes identified experimentally and doesn't have many nodes not experimentally identified. The complete list of 100 genes in the overlap and further statistical analysis are provided in Tables S1 and S3. The dotted line represents the statistically significant limit.

(B) Major direct interaction network of the 248 genes significantly downregulated after transfection of HepG2 cells with miR-24 mimics. Nodes with at least five interactions (including autoregulation) are highlighted.

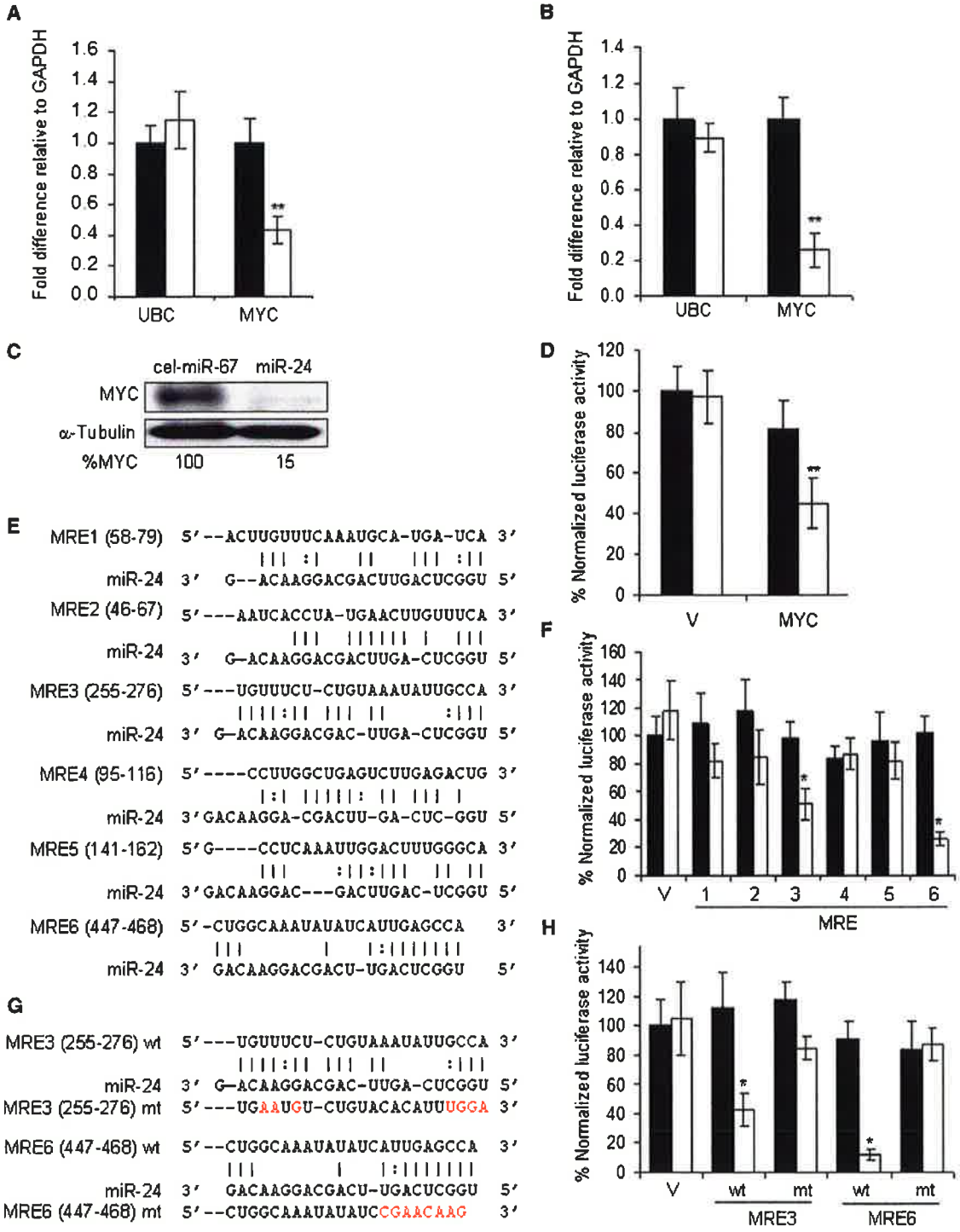


Figure 4. miR-24 Regulates MYC Expression

(A and B) miR-24 overexpression in HepG2 (A) or K562 cells (B) decreases *MYC* mRNA, analyzed by qRT-PCR and normalized to *GAPDH* (black, cel-miR-67; white, miR-24).

(C) MYC protein is decreased in K562 cells upon miR-24 overexpression. Densitometry was used to quantify protein; α-tubulin served as loading control.

(D) miR-24 targets the *MYC* 3′UTR in a luciferase reporter assay. HepG2 cells were transfected with control miRNA (black) or miR-24 (white) mimic for 48 hr and then with *MYC* 3′UTR-luciferase reporter (MYC) or vector (V) for 24 hr.

(E) Predicted binding sites in the *MYC* 3′UTR for miR-24 (MRE1–6) by ma22. The numbers in parentheses correspond to the position in the *MYC* 3′UTR. Perfect matches are indicated by a line; G:U pairs, by a colon.

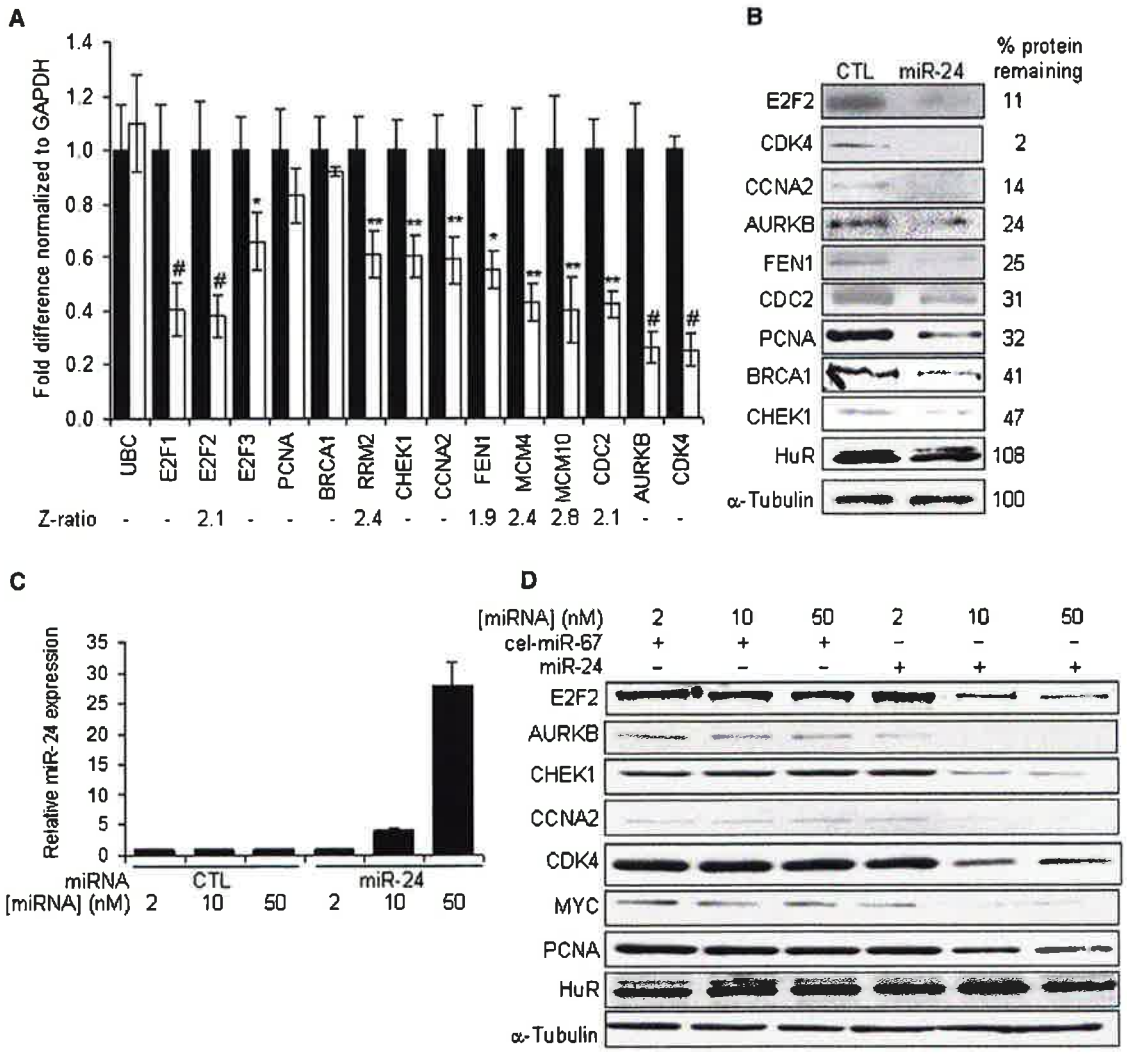


Figure 5. miR-24 Overexpression Downregulates *E2Fs* and Some E2F-Regulated Genes

(A) miR-24 significantly reduces mRNA levels of *E2F1*, *E2F2*, and *E2F3* and of some E2F target genes (*RRM2*, *CHEK1*, *CCNA2*, *FEN1*, *MCM4*, *MCM10*, *CDC2*, and *AURKB*), but not *BRCA1* and *PCNA*. *CDK4*, a key MYC target gene, is also downregulated. HepG2 cells were transfected with miR-24 (white) or control mimics (black) for 48 hr, and *E2F* and their target mRNAs were measured by qRT-PCR. Data normalized to *GAPDH* are expressed relative to control mimic-transfected cells. *UBC* is a control housekeeping gene. The Z ratios refer to the significantly downregulated mRNAs in the mRNA microarray in miR-24-overexpressing cells.

(B) Protein expression of miR-24 target genes is substantially reduced in miR-24 mimic-transfected K562 cells 72 hr after transfection, relative to control miRNA-transfected cells. Densitometry was used to quantify protein relative to α-tubulin. HuR and α-tubulin are loading controls.

(C) miR-24 levels significantly increased when K562 cells were transfected with 10 or 50 nM miR-24 mimics as measured by qRT-PCR analysis. Expression relative to *U6* snRNA is depicted normalized to control cells.

(D) A 4-fold increase in miR-24, obtained by transfecting 10 nM miR-24 mimic, reduces target proteins. Cell lysates of K562 cells, obtained 72 hr after transfection with indicated concentrations of control or miR-24 mimics, were analyzed by immunoblot. α-tubulin and HuR are loading controls.

Error bars represent mean ± SD from three independent experiments. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{\#}p < 0.005$.

248 candidate genes that might be either directly or indirectly regulated by miR-24. Of these downregulated genes, 40% are predicted miR-24-regulated genes by TargetScan, and 53% have a 3′UTR hexamer sequence complementary to the miR-24 seed, suggesting that a large proportion of miR-24-downregulated genes may be direct targets.

To make sense of the set of 248 genes downregulated by miR-24 overexpression, we used bioinformatics to identify

(F–H) miR-24 regulates MRE3 and MRE6 by luciferase reporter assay. HepG2 cells were cotransfected with cel-miR-67 (black) or miR-24 (white) mimics and luciferase reporters containing the wild-type (wt) MRE1-6 in (E) and (F) or mutated (mt) MRE3 and MRE6 in (G) and (H) or vector (V). Luciferase activity was measured 48 hr after transfection. In (G), red letters denote point mutations that disrupt base pairing.

Mean ± SD, normalized to vector control, of three independent experiments is shown. $^{*}p < 0.05$, $^{**}p < 0.01$.

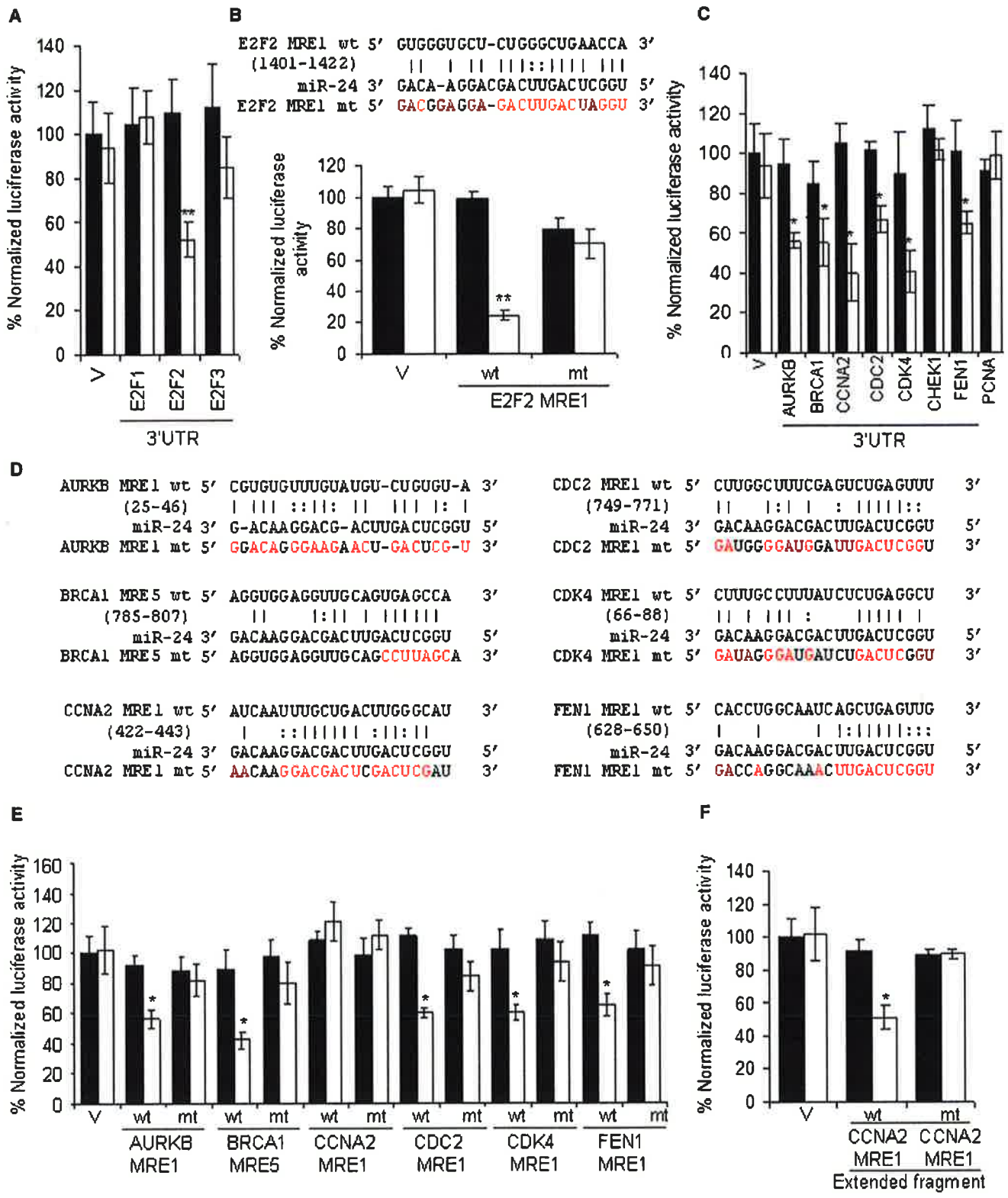


Figure 6. *E2F2* and Multiple E2F Target Genes Are Direct Targets of miR-24, Recognized by Seedless MREs

(A) miR-24 silences the expression of luciferase genes engineered with the 3′UTR of *E2F2*, but not with *E2F1* or *E2F3* 3′UTRs, suggesting that *E2F2* is a direct miR-24 target but *E2F1* and *E2F3* are downregulated indirectly. Luciferase assays were performed in HepG2 cells overexpressing miR-24 (white) or control mimics (black).

overrepresented processes and direct interacting protein networks within this gene set. This type of analysis, which surprisingly does not seem to have been previously applied to understanding miRNA regulation, led to the hypothesis that miR-24 might regulate cell-cycle progression during postmitotic differentiation by targeting *MYC* and/or *E2F2*, given that they constituted nodes of the major interaction network of the downregulated gene set. Both *MYC* and *E2F2* are directly regulated by miR-24, but neither of these genes is a predicted miR-24 target. *MYC*, which has a 3′UTR hexamer seed sequence, is regulated both by a seed-containing MRE and a noncanonical seedless MRE. *E2F2* lacks any miR-24 seed match. However, *E2F2* turned out to be the key gene for miR-24 inhibition of the cell cycle because overexpressing miR-24-insensitive *E2F2* completely restored proliferation.

The GO analysis of miR-24-downregulated genes also suggested that miR-24 might regulate DNA repair. We recently verified this prediction by showing that overexpression of miR-24 enhances sensitivity to DNA damage (Lal et al., 2009). The key miR-24 target for this biological effect is *H2AFX*, which has two seed-bearing predicted MREs.

An unbiased analysis, which did not filter out genes whose 3′UTR lack seed binding sites, was critical for enabling us to identify *E2F2* as the key miR-24 target gene for cell-cycle regulation. In addition to *MYC* and *E2F2*, we found five other miR-24-downregulated genes whose 3′UTR was inhibited by miR-24 through seedless MREs. These genes (*AURKB*, *CCNA2*, *CDK4*, *CDC2*, and *FEN1*) are also transcriptionally regulated by E2Fs or MYC and play crucial roles in cell-cycle progression. Our results suggest that, in addition to genes containing miR-24 perfect seed matches, seedless MREs are also important. Indeed, seedless MREs are critical for miR-24 function because the antiproliferative effect of miR-24 can be recapitulated by silencing or obliterated by overexpressing the seedless *E2F2* gene. However, the importance of recognition of seedless versus seed-bearing MREs could vary between miRNAs. An assessment of this question could be determined by experimental testing of a large set of randomly chosen genes, whose protein or mRNA is downregulated by miRNA overexpression or increased by miRNA inhibition. In addition to seedless 3′UTR MREs, we previously identified a coding region miR-24 MRE in *p16INK4A* (Lal et al., 2008). Other recent studies also identified coding region MREs (Duursma et al., 2008; Tay et al., 2008). Taken together, these results suggest that target gene identification might be improved by not disregarding noncanonical MREs.

miR-24 directly regulates not only critical nodes of the interactome of cell-cycle regulatory genes, but also genes downstream of these nodes. This multitiered gene regulation may guarantee that cell-cycle arrest is not easily evaded. In fact, we have preliminary data suggesting that miR-24 may directly regulate many additional periodic genes, including others that lack a canonical seed-bearing MRE. Because expression of many of these genes is suppressed in nondividing cells, we were careful to show that miR-24-mediated gene suppression occurs before miR-24-transduced cells have stopped dividing (Figure S4), so their downregulation is a cause, not consequence, of cell-cycle arrest.

MYC and E2F regulate progression through G1. Regulating the transition to S phase may be the major site of miR-24 action because miR-24-treated cells accumulate in G1. Because MYC and E2F2 promote each other's transcription, miR-24 may prevent the reciprocal activation of these genes by regulating both of them. The dramatic downregulation of both proteins in miR-24-overexpressing cells could, therefore, be a combined effect of posttranscriptional and transcriptional regulation. Other miR-24 targets that are also transcriptionally regulated by MYC or E2F2 are implicated in controlling progression through G1, the G1/S checkpoint, S, and G2/M. For example, overexpressing miR-24 downregulated the mRNA and protein levels of the E2F-regulated genes, *CCNA2* and *CDC2*, which act together to promote G1/S and G2/M transitions. CCNA2 binds to and activates CDC2, thereby promoting G1/S and G2/M transition. Although the mRNA for *CDK6*, which regulates G1-to-S transition, was not significantly changed in our microarrays, we previously showed that *CDK6* is directly regulated by miR-24 (Lal et al., 2008). *CDK6* may be an example of a target gene regulated primarily by translational inhibition. *p16INK4A* is another direct target of miR-24 that is translationally regulated by miR-24 and, therefore, not downregulated in the microarrays (Lal et al., 2008). In addition to genes, which act at the G1/S transition, cells transfected with miR-24 mimics also have decreased expression of genes that principally act at other phases of the cell cycle. Important genes required for DNA replication in S phase were also downregulated by miR-24, including MCM4 and MCM10 in the prereplication complex; RRM2, a ribonucleotide reductase that catalyzes deoxyribonucleotide synthesis from ribonucleotides; PCNA, which forms a moving platform to recruit replication enzymes to the replication fork; and FEN1, a flap endonuclease involved in rejoining Okasaki fragments. Other downregulated genes act principally to facilitate mitosis, including AURKB and CCNB1. Therefore, miR-24 may put the breaks on cell division at multiple steps in cell-cycle progression.

(B) miR-24 downregulates luciferase activity of a reporter gene containing wild-type (wt) *E2F2* MRE1. Mutations in the miR-24 pairing residues (mt) rescue luciferase expression (sequences and luciferase assays for candidate *E2F2* WT 3′UTR MREs are shown in Figure S3).
(C) miR-24 targets the 3′UTR of E2F-regulated genes (*AURKB*, *BRCA1*, *CCNA2*, *CDC2*, and *FEN1*) and *CDK4*, a MYC-regulated gene. *CHEK1* and *PCNA* 3′UTRs are not regulated by miR-24. HepG2 cells were cotransfected with a luciferase reporter containing the 3′UTR of the indicated gene and control miRNA (black) or synthetic miR-24 (white) for 48 hr. Expression of the unmodified luciferase vector (V) is unchanged by miR-24.
(D) Predicted binding sites in the 3′UTR of genes whose 3′UTR was repressed by miR-24 in (C) and binding site mutations tested (indicated in red).
(E) Expression of reporter genes containing wild-type (wt) *AURKB* MRE1, *BRCA1* MRE5, *CDC2* MRE1, *CDK4* MRE1, and *FEN1* MRE1 is significantly reduced upon cotransfection of HepG2 cells with miR-24 mimics (white) and not the control mimic (black). Mutations in the miR-24 pairing residues (mt) rescue luciferase expression (sequences and luciferase assays for all tested wt MREs for these genes are shown in Figures S6 and S7). The *CCNA2* MRE1 is not regulated.
(F) However, miR-24 regulates a 181 nt region containing the *CCNA2* MRE1 in the luciferase vector. Mutations in the binding residues of *CCNA2* MRE1 within the extended sequence restore luciferase activity.
Error bars in (A), (C), and (E) represent mean ± SD from three independent experiments. $^{**}p < 0.01$, $^{*}p < 0.05$.

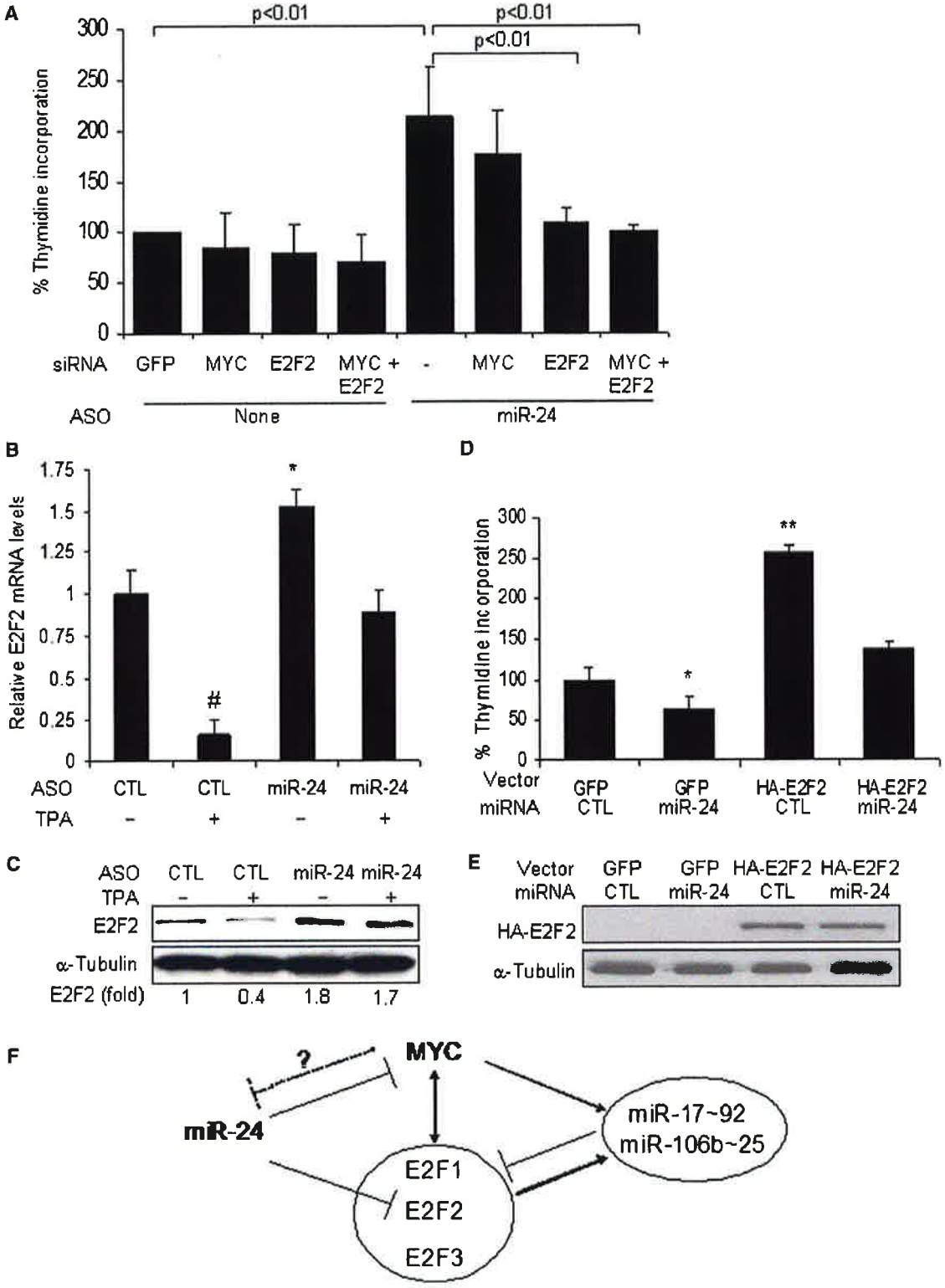


Figure 7. *E2F2* Is a Key miR-24 Target Gene

(A) Increased cell proliferation from antagonizing miR-24 in K562 cells is blocked by siRNA-mediated knockdown of *E2F2*, but not *MYC*. Knockdown is shown by immunoblot (Figure S5). K562 cells were cotransfected with or without miR-24 ASO plus control siRNA or siRNAs targeting *E2F2* and/or *MYC*. The rate of cellular proliferation was determined 72 hr later by thymidine incorporation. Error bars represent mean ± SD from three independent experiments.

Silencing the genes mentioned above would be expected to inhibit cell division. However, suppressing other miR-24-downregulated genes would promote cell-cycle progression, especially in the context of DNA damage. These genes include CHEK1, which participates in the G2/M checkpoint and is activated by ATR in response to unresolved DNA damage, and BRCA1, which is in a surveillance complex that activates double-strand break repair. Prominent in the downregulated gene interaction network are the cyclin D inhibitor *CDKN1B* (*p27KIP1*) and *VHL*, a tumor suppressor protein. In addition, *p16INK4A*, a CDK inhibitor, is a validated direct miR-24 target (Lal et al., 2008). Thus, the role of miR-24 in regulating the cell cycle may be complex. If cells are unable to exit G1 and replicate their DNA, it may be economical to suppress the inhibitory genes that guard the genome from propagating damaged DNA. However, in some contexts, depending on the transcripts expressed in a particular cell, miR-24 might actually promote cell proliferation by suppressing these cell-cycle inhibitory genes. In fact, inhibiting miR-24 decreases proliferation of A549 lung cancer cells but has the opposite effect on HeLa cells (Cheng et al., 2005).

miR-24 is most highly expressed in G1. This is consistent with our finding that miR-24 regulates the G1/S transition. The E2F family of transcription factors regulates progression through this checkpoint. It therefore makes sense that miR-24 acts, in large part, by directly targeting *E2F2* (and thereby indirectly suppressing *E2F1* and *E2F3* expression). The pattern of miR-24 expression is consistent with the known cell-cycle variation of E2F family members (Sears et al., 1997). When miR-24 is high in G1, E2F1 and E2F2 are low; the E2F family begins to be expressed in late G1 and peaks in S phase when miR-24 is turned down. The E2Fs continue to be expressed in G2 and M (where they also have important functions) when miR-24 levels remain low.

miR-24 is not the only miRNA that regulates the cell cycle and targets the *E2F* family (Figure 7F). For instance, the miR-17∼92 cluster directly downregulates the *E2F* family (O'Donnell et al., 2005; Petrocca et al., 2008). However, unlike miR-24, whose expression varies inversely with E2F expression, miR-17∼92 appears to be expressed uniformly except in quiescent cells. Moreover, E2F2 downregulation should antagonize the dominant effect of these miRNAs to promote cell proliferation. Thus, E2F downregulation is likely not a defining effect of miR-17∼92 but, rather, a secondary effect that fine-tunes its major proliferative effect. Because miR-24 suppresses *MYC* and *E2F* expression and both MYC and the E2F family activate miR-17∼92 and miR-106b∼25 transcription (O'Donnell et al., 2005; Petrocca et al., 2008), miR-24 also likely inhibits proliferation by indirectly suppressing transcription of these cell-cycle-promoting miRNAs. A recent paper also suggests another layer of complexity to the miR-24, miR-17∼92, MYC, and E2F network (Gao et al., 2009). MYC suppresses the transcription of miR-23b, which is encoded with miR-24. Although one recent study suggests that miR-24 and miR-23b are independently transcribed (Sun et al., 2009), MYC might also regulate miR-24 transcription. It is worth noting that *E2F1* mRNA and protein do not correlate during the cell cycle, consistent with posttranscriptional regulation by miRNAs (O'Donnell et al., 2005). Because *E2F2* mRNA has kinetics similar to *E2F1* (Sears et al., 1997), miRNA-dependent regulation of translation may be operating, possibly for all E2Fs that promote G1/S transition.

The integrated effect of miR-24 on a highly interacting set of key genes acts as a switch to stop cell division, rather than as a fine-tuning rheostat. It will be interesting to understand how expression of these two miR-24 gene clusters is regulated and to understand the function of the clustered miRNAs (miR-23 and miR-27). The only other miRNAs consistently upregulated during terminal differentiation are miR-22 and miR-125a (a mammalian ortholog of lin-4) (Lal et al., 2009). There are suggestions in the literature that these genes might also regulate important pathways of cell differentiation (Choong et al., 2007; Wu and Belasco, 2005).

miR-24 directly regulates both cell proliferation and DNA repair. Enhancing miR-24 function in cancer cells by introducing miR-24 mimics might be an attractive therapeutic, given that it could potentially block dysregulated cell proliferation and also sensitize cancer cells to DNA damage from chemo- and radiotherapy.

EXPERIMENTAL PROCEDURES

GO Analysis

The GO project provides structured controlled vocabularies, or ontologies, to describe genes relative to their biological processes. Each biological process consists of a series of events achieved by one or more molecular functions. The ontologies are stored in directed acyclic graphs in which each node represents a biological process and each subsequent node corresponds to a more specialized term. Overrepresented GO biological processes were determined using a MetaCore Analytical Suite (GeneGo Inc., St. Joseph, MI, http://www.genego.com), which utilizes the hypergeometric distribution to calculate the p value for genes showing enrichment in a biological process. The value is equivalent to the probability of a subset of genes from a specific

(B and C) Downregulation of *E2F2* mRNA (B) and protein (C) during TPA-mediated differentiation of K562 cells to megakaryocytes is mediated by miR-24 and can be completely inhibited by antagonizing miR-24. K562 cells were transfected with miR-24 or a control (CTL) ASO for 72 hr and then treated with TPA for 6 hr. mRNA was assessed by qRT-PCR normalized to *GAPDH* and normalized to control cells transfected with CTL ASO. E2F2 protein was quantified by densitometry and normalized to α-tubulin.

(D and E) Transfection of K562 cells with a miR-24 mimic reduces cell proliferation, which can be rescued by expressing miR-24-insensitive *E2F2* lacking the 3′UTR. K562 cells were cotransfected with a vector expressing HA-tagged E2F2 or GFP and miR-24 or cel-miR-67 (CTL) mimics for 72 hr before measuring thymidine uptake.

(E) Immunoblot probed for HA tag. Error bars represent mean ± SD from three independent experiments. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{\#}p < 0.005$.

(F) Model of the miR-24, miR-17∼92, MYC, and E2F network of cell-cycle regulators. Here, we show that miR-24 directly suppresses expression of *MYC* and *E2F2* (and indirectly suppresses *E2F1* and *E2F3*) and thereby regulates the G1/S transition. Expression of the opposing miRNAs encoded by the miR-17∼92 and miR-106b∼25 clusters that promote cell proliferation is transcriptionally activated by the same transcription factors that miR-24 suppresses (O'Donnell et al., 2005; Petrocca et al., 2008). Therefore, miR-24 would be predicted to reduce expression of the proliferation-promoting miRNA clusters indirectly. These miRNAs also knock down the *E2F* genes but probably to fine-tune their proliferative effect. MYC may also suppress miR-24 transcription (Gao et al., 2009).

experiment (i.e., miR-24 overexpression) to arise by chance given the total number of genes associated with the biological process.

Network Analysis

We developed a graphical representation of the molecular relationships between proteins from the 248 genes downregulated in miR-24-overexpressing cells. Network analysis was performed using the MetaCore Analytical Suite, and visualization was performed using Ingenuity Pathways Analysis (Ingenuity Inc., http://www.ingenuity.com). Proteins are represented as nodes, and the biological relationship between two nodes is represented as an edge. Functional and pathway analysis was based on support for direct interaction edges in the networks from at least one reference from the literature, a textbook, or from canonical information stored in the network generation software database, manually curated for experimentally verified human protein-protein interactions and protein-DNA interactions.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, nine figures, and five tables and can be found with this article online at http://www.cell.com/molecular-cell/supplemental/S1097-2765(09)00600-5.

ACKNOWLEDGMENTS

This work was supported, in part, by National Institutes of Health (NIH) AI070302 and a GSK-IDI Alliance grant (J.L.); the NIA-IRP, NIH (K.G.B., M.G.); the Harry Oppenheimer Memorial Trust (W.H.); a GSK-IDI Alliance fellowship (F.N.); and the Harvard Center for AIDS Research (N.Y.). We thank N. Dyson (Harvard Medical School) for the HA-E2F2 expression plasmid, Ray McGovern for programming support, and Lieberman laboratory members for useful discussions.

Received: December 24, 2008
Revised: August 10, 2009
Accepted: August 25, 2009
Published: September 10, 2009

REFERENCES

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64–71.

Bar-Joseph, Z., Siegfried, Z., Brandeis, M., Brors, B., Lu, Y., Eils, R., Dynlacht, B.D., and Simon, I. (2008). Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. Proc. Natl. Acad. Sci. USA *105*, 955–960.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Bracken, A.P., Ciro, M., Cocito, A., and Helin, K. (2004). E2F target genes: unraveling the biology. Trends Biochem. Sci. *29*, 409–417.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M.V., Visone, R., Sever, N.I., Fabbri, M., et al. (2005). A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N. Engl. J. Med. *353*, 1793–1801.

Chang, T.C., Wentzel, E.A., Kent, O.A., Ramachandran, K., Mullendore, M., Lee, K.H., Feldmann, G., Yamakuchi, M., Ferlito, M., Lowenstein, C.J., et al. (2007). Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. Mol. Cell *26*, 745–752.

Cheng, A.M., Byrom, M.W., Shelton, J., and Ford, L.P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. Nucleic Acids Res. *33*, 1290–1297.

Choong, M.L., Yang, H.H., and McNiece, I. (2007). MicroRNA expression profiling during human cord blood-derived CD34 cell erythropoiesis. Exp. Hematol. *35*, 551–564.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat. Struct. Mol. Biol. *13*, 849–851.

Didiano, D., and Hobert, O. (2008). Molecular architecture of a miRNA-regulated 3′ UTR. RNA *14*, 1297–1317.

Diederichs, S., and Haber, D.A. (2007). Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. Cell *131*, 1097–1108.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes Dev. *18*, 504–511.

Duursma, A.M., Kedde, M., Schrier, M., le Sage, C., and Agami, R. (2008). miR-148 targets human DNMT3b protein coding region. RNA *14*, 872–877.

Easow, G., Teleman, A.A., and Cohen, S.M. (2007). Isolation of microRNA targets by miRNP immunopurification. RNA *13*, 1198–1204.

Fernandez, P.C., Frank, S.R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2003). Genomic targets of the human c-Myc protein. Genes Dev. *17*, 1115–1129.

Fukuda, Y., Kawasaki, H., and Taira, K. (2005). Exploration of human miRNA target genes in neuronal differentiation. Nucleic Acids Symp. Ser. (Oxf.) *49*, 341–342.

Gao, P., Tchernyshyov, I., Chang, T.C., Lee, Y.S., Kita, K., Ochi, T., Zeller, K.I., De Marzo, A.M., Van Eyk, J.E., Mendell, J.T., and Dang, C.V. (2009). c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. Nature *458*, 762–765.

Hermeking, H., Rago, C., Schuhmacher, M., Li, Q., Barrett, J.F., Obaya, A.J., O'Connell, B.C., Mateyak, M.K., Tam, W., Kohlhuber, F., et al. (2000). Identification of CDK4 as a target of c-MYC. Proc. Natl. Acad. Sci. USA *97*, 2229–2234.

Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D., and Slack, F.J. (2005). RAS is regulated by the let-7 microRNA family. Cell *120*, 635–647.

Johnson, C.D., Esquela-Kerscher, A., Stefani, G., Byrom, M., Kelnar, K., Ovcharenko, D., Wilson, M., Wang, X., Shelton, J., Shingara, J., et al. (2007). The let-7 microRNA represses cell proliferation pathways in human cells. Cancer Res. *67*, 7713–7722.

Lal, A., Kim, H.H., Abdelmohsen, K., Kuwano, Y., Pullmann, R., Jr., Srikantan, S., Subrahmanyam, R., Martindale, J.L., Yang, X., Ahmed, F., et al. (2008). p16(INK4a) translation suppressed by miR-24. PLoS ONE *3*, e1864.

Lal, A., Pan, Y., Navarro, F., Dykxhoorn, D., Moreau, L., Meire, E., Bentwich, Z., Lieberman, J., and Chowdhury, D. (2009). miR-24-mediated down-regulation of H2AX suppresses DNA repair in terminally differentiated blood cells. Nat. Struct. Mol. Biol. *16*, 492–498.

Lebofsky, R., and Walter, J.C. (2007). New Myc-anisms for DNA replication and tumorigenesis? Cancer Cell *12*, 102–103.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787–798.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773.

Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203–1217.

Neilson, J.R., Zheng, G.X., Burge, C.B., and Sharp, P.A. (2007). Dynamic regulation of miRNA expression in ordered stages of cellular development. Genes Dev. *21*, 578–589.

O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V., and Mendell, J.T. (2005). c-Myc-regulated microRNAs modulate E2F1 expression. Nature *435*, 839–843.

Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. *216*, 671–680.

Petrocca, F., Visone, R., Onelli, M.R., Shah, M.H., Nicoloso, M.S., de Martino, I., Iliopoulos, D., Pilozzi, E., Liu, C.G., Negrini, M., et al. (2008). E2F1-regulated microRNAs impair TGFbeta-dependent cell-cycle arrest and apoptosis in gastric cancer. Cancer Cell *13*, 272–286.

Polager, S., and Ginsberg, D. (2008). E2F - at the crossroads of life and death. Trends Cell Biol. *18*, 528–535.

Sears, R., Ohtani, K., and Nevins, J.R. (1997). Identification of positively and negatively acting elements regulating expression of the E2F2 gene in response to cell growth signals. Mol. Cell. Biol. *17*, 5227–5235.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Sun, Q., Zhang, Y., Yang, G., Chen, X., Cao, G., Wang, J., Sun, Y., Zhang, P., Fan, M., Shao, N., and Yang, X. (2008). Transforming growth factor-beta-regulated miR-24 promotes skeletal muscle differentiation. Nucleic Acids Res. *36*, 2690–2699.

Sun, F., Wang, J., Pan, Q., Yu, Y., Zhang, Y., Wan, Y., Wang, J., Li, X., and Hong, A. (2009). Characterization of function and regulation of miR-24-1 and miR-31. Biochem. Biophys. Res. Commun. *380*, 660–665.

Tay, Y., Zhang, J., Thomson, A.M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature *455*, 1124–1128.

Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K., and Slack, F.J. (2004). The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. Genes Dev. *18*, 132–137.

Vernell, R., Helin, K., and Muller, H. (2003). Identification of target genes of the p16INK4A-pRB-E2F pathway. J. Biol. Chem. *278*, 46124–46137.

Wu, L., and Belasco, J.G. (2005). Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. Mol. Cell. Biol. *25*, 9198–9208.

Zhang, L., Ding, L., Cheung, T.H., Dong, M.Q., Chen, J., Sewell, A.K., Liu, X., Yates, J.R., III, and Han, M. (2007). Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. Mol. Cell *28*, 598–613.